IDENTIFICATION OF AN ANCIENT *BMP4* CIS-REGULATORY ELEMENT USING FISH AND MOUSE

By

Kelly Jane Chandler

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Molecular Physiology and Biophysics

August, 2008

Nashville, Tennessee

Approved:

Professor Maureen A. Gannon

Professor Richard M. O'Brien

Professor Ela W. Knapik

Professor E. Michelle Southard-Smith

Professor Linda J. Sealy

ACKNOWLEDGEMENTS

A dissertation is not a body of one person's work. It is a body of work lead by one person and supported by many. In this regard, I would like to thank those who have made this body of work possible for me to achieve. I'd like to thank the Department of Molecular Physiology and Biophysics for providing the training and support I have received. I'd also like to thank the Developmental Biology Program for allowing me to partake in their retreats and journal clubs, which were very beneficial to my education. I will miss the smiling face of Yue Hou: a dedicated scientist, caring surrogate mother, and young girl at heart. Thank you for your help and your friendship. To the women scientists who inspired and encouraged me to attend graduate school: Dr. Mary Allen, Dr. Vicki Rosen and Dr. Jane Owens. You have made a difference in my life. I hope to do the same for other young women in the future. A special thank you to my thesis committee: Dr. Maureen Gannon, Dr. Richard O'Brien, Dr. Linda Sealy, Dr. Michelle Southard-Smith and Dr. Ela Knapik. You challenged me, respected me, and encouraged me throughout my time at Vanderbilt and I will take that with me wherever I go. To Dr. Douglas Mortlock, my mentor. Thank you for giving me a position in your laboratory. You always took the time to "talk science" and regardless of how much red ink you put on something I had written, you were always encouraging. Most of all, thank you for your guidance, support, and Bmp4! A big thank you to my family for their love and support through the years. To my parents, thank you for loving me and giving me every opportunity in the world. Thank you for instilling the values of hard work, integrity, and character in

me. Without this, I would not be the person I am today. Mom, thanks for giving me guts and a sense of humor and, of course, for all the countless ways you have supported me throughout graduate school. Marshall and Amanda, thank you for putting up with the evil grad school vacuum that was my life for the past six years. I owe you. Thanks Grandpa for always asking me how school was "coming along" and Moms for always telling me I could do it when I wasn't sure I could. To Cowboy, thank you for teaching me responsibility, unconditional love, and tenacity through dark times. I've got so much to tell you when we meet again. To my daughter, Tennyson, I have arrived at this point in my career by way of few successes and many failures. May you always have the courage to fail and the determination and perseverance to come back from failure. To my soul mate, true love and best friend, Ron. You have changed my life in ways I never thought were possible. I never imagined I'd find you across the bench and I knew choosing you might shadow my abilities for some, but I'd do it the same way a million times over. Thank you for being the man I always dreamed of.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

BACKGROUND AND SIGNIFICANCE

Bone morphogenetic proteins (BMPs) comprise a subfamily of at least twenty distinct proteins in the Transforming Growth Factor-Beta (TGF-β) superfamily of secreted signaling molecules. Initially, BMPs were discovered based on their ability to induce the formation of ectopic bone (Urist 1965). The first Bmps were cloned in 1988 and were designated Bone Morphogenetic Proteins since they were able to induce the bone-forming cascade *in vivo* (Wozney et al. 1988). Although the name implies a bone-specific function for this class of proteins, research over several decades has demonstrated BMPs play dynamic roles across multiple tissues throughout embryonic development (Zhao 2003) (Hogan 1996). *Bmp4*, initially designated *Bmp-2b*, was among the first of the Bmps to be cloned (Wozney et al. 1988). Comparative analysis of the carboxy terminal portion of *BMP4* and *BMP2* (formerly *BMP-2A*) revealed significant conservation (92%) between the highly homologous proteins (Wozney et al. 1988). In addition, BMP4 showed significant conservation (~75%) with the Drosophila decapentaplegic protein (DPP) suggesting BMP4 may be the human homolog of DPP (Wozney et al. 1988). Likewise, DPP was capable of inducing bone formation in an *in vivo* mammlian system (Sampath et al. 1993) suggesting the distantly related proteins can function interchangeably.

Bmp4 Signaling and Regulation

Like all TGF-β superfamily proteins, BMPs are synthesized as precursor proteins with three signature motifs: 1) an amino-terminal signal sequence 2) a propeptide sequence 3) and a mature carboxy terminus. Unlike other TGF-β superfamily members, BMPs are characterized by seven conserved cysteines in the mature region of the protein. BMPs are synthesized as inactive precursor proteins and subsequently form homodimers or heterodimers through a single disulfide linkage (Aono et al. 1995) (Hazama et al. 1995) (Constam and Robertson 1999). BMP4 dimers must be cleaved in the prodomain by endoproteases before the first cysteine in the mature region at a dibasic primary RXR/KR motif and then at a secondary RXXR motif to render the protein active (Cui et al. 1998) (Constam and Robertson 1999). BMP4 precursor protein has been shown to heterodimerize with BMP7 and display significantly increased activity compared to the BMP4 homodimer complex (Aono et al. 1995) (Hazama et al. 1995). The propeptide portion of BMP4 imparts stability on the mature protein and it determines how efficiently the protein is secreted (Constam and Robertson 1999) as well as the protein's activity (Cui et al. 2001). Thus, the mature C-terminal dimer is the biologically active form of the protein.

The biologically active form of BMP4 signals through type I/ type II serine threonine kinase receptors (FIGURE 1.1). Multiple type I and type II receptors have been described and the combination of these receptors imparts ligand binding specificity. However, BMP4 is believed to utilize BMPRII (type I receptor) complexed with BMPRIA (ALK3), BMPRIB (ALK6), or ActRIA (ALK2) (type II

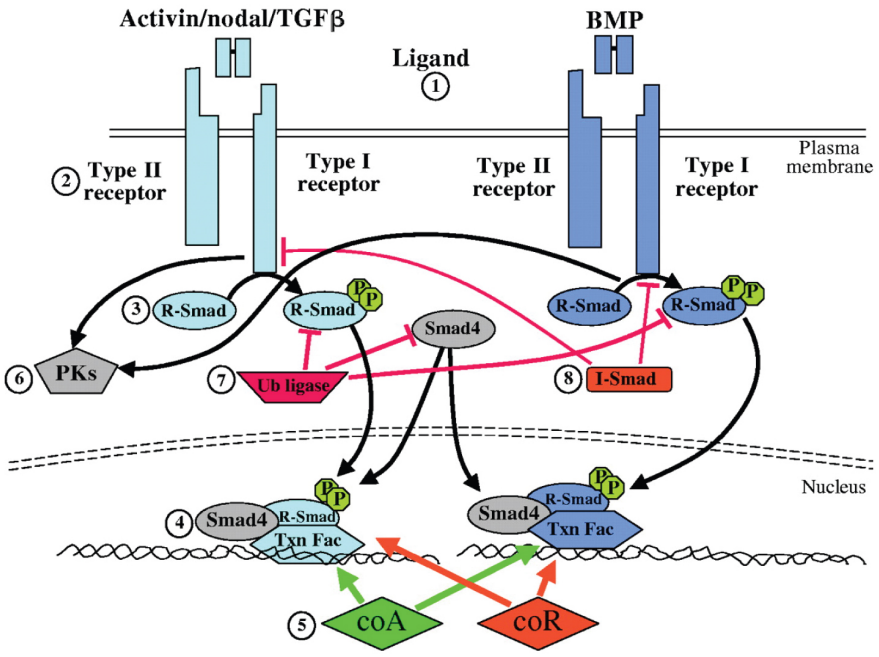

Figure 1.1. Bmp signaling is mediated by serine/threonine kinase receptors and intracellular Smad molecules. Upon binding Type I and Type II serine threonine kinase receptors, Bmp and its cognate receptors form a heterotetrameric signaling complex. In turn, receptor-Smad (R-Smad) molecules become phosphorylated resulting in the nuclear translocation of R-Smad. R-Smad then forms a complex with Smad4 and a transcription factor that is capable of binding the DNA target sequence to repress or activate target gene expression. Adapted from (Whitman and Raftery 2005)

receptors) (Mishina 2003). Upon binding with the receptor complex, BMP4 initiates a signaling cascade whereby the type II receptors phosphorylate the type I receptors which, in turn, phosphorylate intracellular signaling molecules known as Smads (Mishina 2003). BMP4 signaling is believed to phosphorylate Smad1, 5, and 8 which then complex with a Co-Smad (Smad4) allowing translocation into the nucleus and regulation of Bmp4 target genes (Nakayama et al. 2000).

BMP4 signaling is regulated both extracellulary and intracellularly. Once a homodimeric or heterodimeric BMP4 complex is activated by proteolytic cleavage, it binds to a type I/type II serine/threonine kinase receptor to form a heterotetrameric signaling complex (Chen et al. 2004). However, secreted inhibitors such as noggin bind BMP4 and prevent interaction with its receptors (Zimmerman et al. 1996) (Groppe et al. 2002). There are seven secreted BMP antagonists described to date and each has been shown to inhibit BMP4 (Balemans and Van Hul 2002). An alternative extracellular antagonist to BMP4 is a pseudoreceptor called BAMBI/Nma. Experiments in Xenopus have shown that BAMBI can bind to BMP4 and prevent the propogation of a signaling cascade since the pseudoreceptor does not have an intracellular kinase domain (Onichtchouk et al. 1999).

BMP4 signaling is also regulated intracellularly. Within the cytoplasm, inhibitory Smads can bind to activated type I/type II serine threonine kinase receptors and become activated much like Smads 1, 5 and 8. However, unlike these Smads, inhibitory Smads are unable to bind DNA and regulate BMP4

target genes creating a block in downstream signaling (Canalis et al. 2003). Likewise, the Ski protein is capable of binding Smad 1 and 5 as well as Smad 4, but is not capable of regulating target genes in the nucleus (Canalis et al. 2003). Finally, Smurf proteins interact with Smad 1 and 5 promoting the degradation of this complex by the ubiquitin-proteasome pathway (Canalis et al. 2003). Each of these intracellular antagonists negatively regulates BMP4 signaling. Thus, BMP4 signaling can be modulated both within and outside the cell.

Bmp4 Transcriptional Regulation

Despite the significant volume of research on Bmp4 since its discovery nearly twenty years ago, little has been published regarding the transcriptional regulation of *Bmp4*. A genomic clone containing the mouse *Bmp4* gene was isolated and characterized in the early 1990's (Kurihara et al. 1993). In this study, Kurihara *et al.* showed mouse *Bmp4* contained five exons, with the last two exons containing coding regions. In addition, *Bmp4* contained alternative transcriptional start sites in exon I and exon II (promoter 1A, promoter 1B) and each lacked a TATA box (Kurihara et al. 1993). Two years later, another research group corroborated the findings of Kurihara *et al.* and utilized *in vitro* experiments to suggest the alternate transcripts were tissue-specific (Feng et al. 1995). Similar to mouse *Bmp4*, human *BMP4* genomic structure was characterized and found to have cell type-specific alternate transcripts derived from five exons (van den Wijngaard et al. 1996) (Van den Wijngaard et al. 1999). In addition, the transcriptional start sites of two functional TATA-less *BMP4* promoters were

mapped and displayed distinct activity in different cell lines suggesting *BMP4* regulation is complex (Van den Wijngaard et al. 1999). Likewise, investigation of mouse *Bmp4* regulation in an osteoblastic cell line suggested the TATA-less promoter 1A was primarily used (Ebara et al. 1997). Analysis of human BMP4 promoter activity in lung epithelial cell lines showed distinction between the use of promoter 1A (primarily used) and promoter 1B in a cell type-specific manner as well (Zhu et al. 2004). Interestingly, Thompson *et al.* discovered a third promoter within intron 2 using multiple independent methods such as 5' rapid amplification of cDNA ends (RACE) and RNase protection assays (RPA) in an immortalized mouse otocyst cell line (Thompson et al. 2003). An independent group, who showed that all three promoters were active in mouse spermatagonia cells, further validated these results (Baleato et al. 2005). In general, both mouse and human *Bmp4* have been shown to contain five exons with an alternative transcript involving exon 1 as depicted by the ECR browser (FIGURE 1.2) (Ovcharenko et al. 2004). These initial studies provided insight into the structure of *Bmp4*; however, the transcriptional assays were clearly limited to cell lines and not a global view of *Bmp4* regulation.

Since *Bmp4* is expressed in a dynamic, spatiotemporal-specific manner throughout development (Jones et al. 1991), it is necessary to assay *Bmp4* transcriptional activity in the developing embryo to obtain a global view of *Bmp4* regulation. Minimal *Bmp4* promoter fragments encompassing promoter 1A, but not promoter 1B or the intron 2 promoter, were tested in transgenic mice for

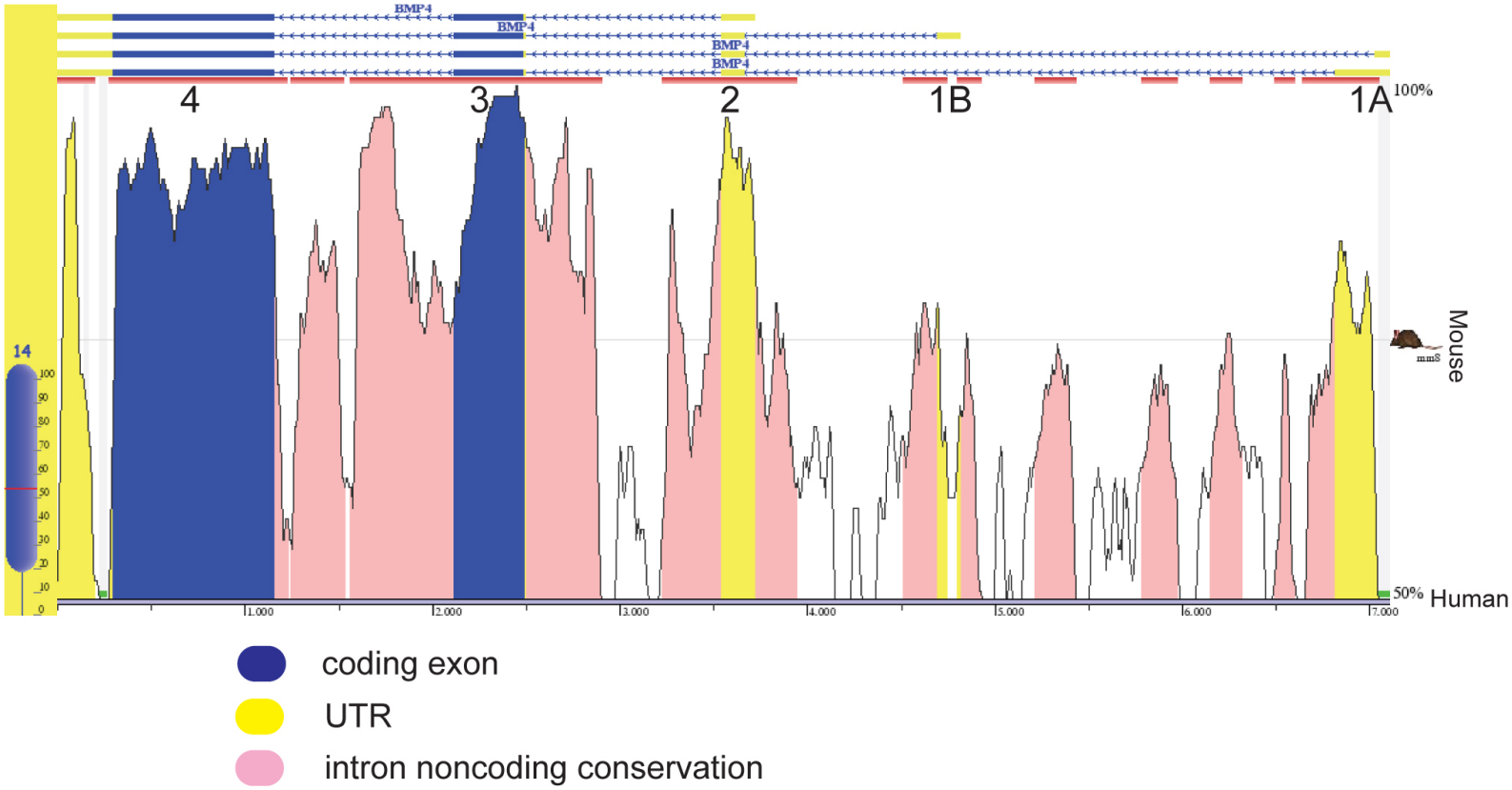

Figure 1.2. *BMP4* exon structure. Shown here is the exon structure of human BMP4 on the negative strand of Chromosome 14 as depicted by the ECR Browser (Ovcharenko et al. 2004). Note, the human segment is being compared to the syntenic region in mouse highlighting the conservation of Bmp4's exon structure. Mouse and human BMP4 are comprised of five exons, four of which are transcribed. Exon 1A and 1B are alternately transcribed followed by Exon 2-4. Exon 3 and 4 are coding exons (blue bars), while Exon 1A, 1B and 2 are not translated. Adapted from http://ecrbrowser.dcode.org.

reporter activity (Feng et al. 2002) (Zhang et al. 2002) and compared to endogenous *Bmp4* expression as shown in the *Bmp4*^{lacZneo} knock-in reporter mouse (FIGURE 1.3) (Lawson et al. 1999). Of the three minimal promoter fragments tested, two showed reporter activity *in vivo*, whereas the smallest promoter fragment failed to exhibit reporter activity in two independent transgenic mouse lines. Upon analyzing embryos at 11.5 days post coitus (dpc), the only expression pattern driven by the 2.4kb and 1.1kb promoter fragments was a segmental pattern along the dorsal region of the embryo (FIGURE 1.3) (Feng et al. 2002). When compared to the *Bmp4*^{lacZneo} knock-in reporter mouse, it is clear that this is an endogenous expression pattern. However, there are numerous sites of expression present in the *Bmp4*^{lacZneo} knock-in reporter mouse at 11.5 dpc that are clearly not driven by either minimal promoter fragments tested (FIGURE 1.3), suggesting most regulatory elements critical for early developmental expression of *Bmp4* reside beyond the regions tested. One out of three reported endogenous promoters were incorporated into the minimal promoter fragments tested suggesting the remaining two promoters may be critical for early *Bmp4* expression. Later in mouse development at 16 dpc, two of the three minimal promoter fragments directed reporter expression in the hair shaft and distal hair matrix, but not in the dermal papilla suggesting regulatory elements for the former expression patterns are present in the minimal promoter fragments, whereas a dermal papilla enhancer most likely resides beyond this minimal region tested (Zhang et al. 2002). Likewise, the two larger minimal promoter fragments directed reporter expression in the tooth ameloblasts,

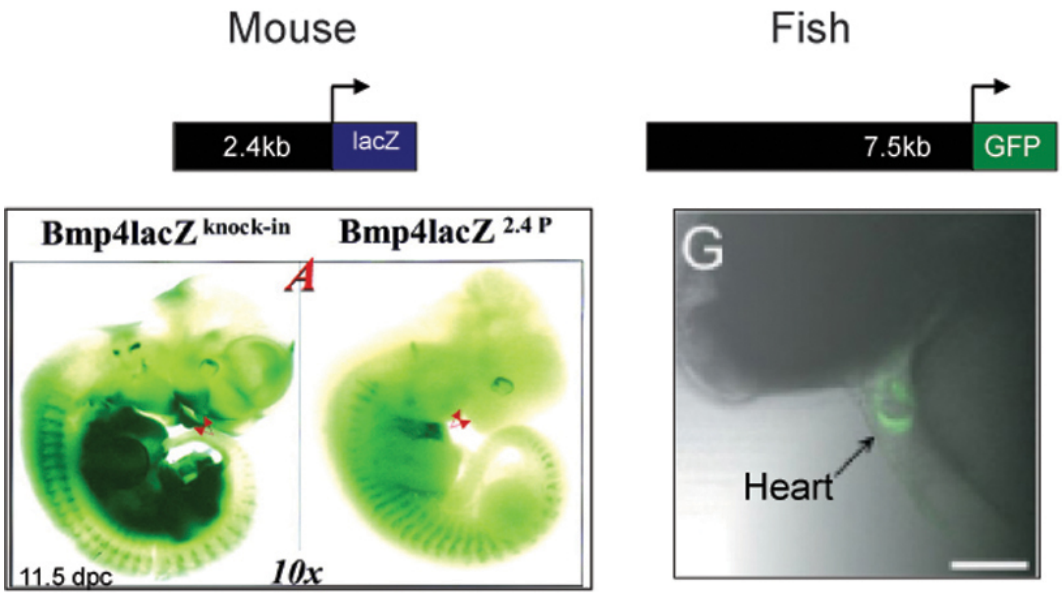

Figure 1.3. *Bmp4* promoter fragments are not sufficient to reproduce all known sites of endogenous expression in mouse and fish. A 2.4 kb *Bmp4* promoter fragment tested in mouse directed expression in a segmental pattern along the dorsal region of the embryo similar to the knock-in line (*Bmp4lacZ* $^{knock\text{-}in}$), but failed to direct many known sites of *Bmp4* expression such as dorsal retina, inner ear, and forebrain. Likewise, a 7.5 kb Bmp4 promoter fragment directed reporter expression in the zebrafish heart. However, the reporter expression failed to mimic endogenous expression suggesting *cis*-regulatory elements reside beyond the fragment tested. Adapted from (Feng et al. 2002) and (Shentu et al. 2003).

tongue, nasal cartilages, bone and salivary glands of newborn mice but not in numerous other known sites of *Bmp4* expression (Zhang et al. 2002). Taken together, these studies suggest that many critical regulatory regions necessary for directing reporter expression in an endogenous manner reside beyond the limits of the largest fragment tested (2.4kb) and may be located in more distant 5', 3' or intronic regions of *Bmp4* (FIGURE 1.3). Likewise, zebrafish promoter/ reporter fragments were used to generate stable transgenic lines to assay *Bmp4* promoter activity (Shentu et al. 2003). While *Bmp4* is expressed in a tissue-specific manner throughout zebrafish development, a 7.5 kb promoter fragment failed to direct any expression patterns similar to *Bmp4* (FIGURE 1.3) (Shentu et al. 2003). This study strongly indicates *Bmp4* expression in zebrafish is complex and *cis*-regulatory elements necessary for proper *Bmp4* expression reside beyond the 7.5 kb fragment that was tested (Shentu et al. 2003). Identification of distant *Bmp4* tissue-specific enhancers will be imperative for understanding how *Bmp4* is activated in a spatiotemporal manner throughout development and postnatally.

Bmp4 Transcriptional Regulation is Complex

Although gene regulation can occur by processes other than transcription (RNA splicing, RNA stability, protein modifications), the focus here is on transcriptional regulation. Despite a lack of enhancer mapping data for *Bmp4*, a significant amount of research has shown that *Bmp4* expression is complex. *In situ* hybridization studies in mouse have provided an initial glimpse of the

complex spatiotemporal regulation of *Bmp4*. These studies revealed *Bmp4* transcripts are present very early in mouse development at 6.5 dpc and persist throughout development in a tissue-specific manner (Jones et al. 1991). Extraembryonic expression is detected in the allantois and amnion at 7.5 dpc followed by localized transcripts present in the mesoderm and endoderm of the posterior primitive streak, ventral mesoderm, myoepicardium of the heart, allantois and amnion one day later (8.5 dpc) (Jones et al. 1991). Just one half day later, *Bmp4* expression is noted in the diencephalon, otic vesicles, first branchial arch, atrioventricular canal of the heart, mesoderm surrounding the gut and lung, somatic and splanchnic mesoderm, and mesenchyme in the flank adjacent to the forelimb bud at 9.0 dpc (Jones et al. 1991). By 10.5 dpc, *Bmp4* is restricted to the floorplate of the diencephalon, nasal pit ectoderm, distal ectoderm of the facial processes, mesenchyme surrounding the gut, myocardial layer of the truncus arteriosus in the heart, apical ectodermal ridge (AER) of the limbs, and limb bud mesenchyme (Jones et al. 1991). Precise regulation of *Bmp4* is evident by the lack of expression in the atrioventricular canal of the heart, a structure that is still present at 10.5 dpc and where *Bmp4* is specifically expressed just one day and a half earlier, and presence of *Bmp4* in the truncus arteriosus of the heart (Jones et al. 1991). This theme continued with *Bmp4* expression present in the same organs, but in different tissues of those organs as development proceeds (11.5-17.5 dpc) (Jones et al. 1991). The sensitivity of *in situ* hybridization is somewhat limited, therefore subsequent creation of the $Bmp4^{lacZneo}$ knock-in reporter mouse allowed for a more detailed analysis of

Bmp4 expression and revealed the onset of expression is actually three days prior (3.5 dpc) to previous reports (6.5 dpc) in mouse blastocysts (Lawson et al. 1999). In addition, X-Gal staining of *Bmp4*^{lacZneo} heterozygous embryos revealed expression in extraembryonic ectoderm at 6.0 dpc versus the low levels of expression detected by *in situ* hybridization at the same time point (Lawson et al. 1999) (Jones et al. 1991). A combination of *in situ* hybridization, RT-PCR, and reporter expression of *Bmp4*^{lacZneo} heterozygous mice over the years have demonstrated *Bmp4* expression in numerous tissue-specific sites throughout mouse development (TABLE 1.1).

Mesoderm Development

Bmp4 has been shown to play a critical role in mesoderm development. Mesoderm tissue is one of three component germ layers of the developing embryo and it arises from the epiblast to form the primitive streak during the onset of gastrulation at 6.5 dpc in mouse development (FIGURE 1.4) (Lu et al. 2001). The formation of the primitive streak is the first definitive sign that gastrulation has begun and the initial appearance of the primitive streak marks the future posterior of the embryo, thereby establishing the anteroposterior axis. Initial fate mapping experiments using single cell injections of horseradish peroxidase were performed in pre-gastrulating mouse embryos to show that cells from most portions of the epiblast could differentiate into both extraembryonic and embryonic mesoderm (Lawson et al. 1991). Subsequent fate mapping experiments in mouse using transplanted cells that constitutively express *lacZ*

have indicated the primary population of cells that migrate from the epiblast through the primitive streak eventually form extraembryonic mesoderm, while cells that are destined to form embryonic mesoderm continue to reside in the epiblast at this stage (Parameswaran and Tam 1995) (Kinder et al. 1999). Extraembryonic mesoderm differentiates to form the amnion, allantois, and chorion by 7.5 dpc (FIGURE 1.4) (Lu et al. 2001). Together, these three structures form the mesodermal lining of the exocoelomic cavity. Later during gastrulation, epiblast cells migrate through the primitive streak to give rise to embryonic lateral plate, paraxial, heart, and cranial mesoderm (Kinder et al. 1999). Interestingly, cells that migrate through the posterior portion of the primitive streak give rise to extraembryonic mesoderm whereas cells that pass through the anterior portion of the primitive streak give rise to embryonic mesoderm (Kinder et al. 1999). This suggests there is a correlation between the location of cell migration in the primitive streak and the cell fate. However, there is no correlation between the location of a precursor cell in the epiblast and the eventual fate of the cell (Kinder et al. 1999) (Lawson et al. 1991).

Multiple distinct tissues arise from extraembryonic and embryonic mesoderm that, in turn, both originate from the epiblast portion of the developing mouse embryo (FIGURE 1.5) (Lu et al. 2001). As organogenesis proceeds, extraembryonic mesoderm differentiates to form the amnion, chorion, allantois, yolk sac, fibroblasts, capillary epithelium, blood vessels of the umbilical cord, placenta, and hematopoietic precursor cells (Hogan 1994). The primary function

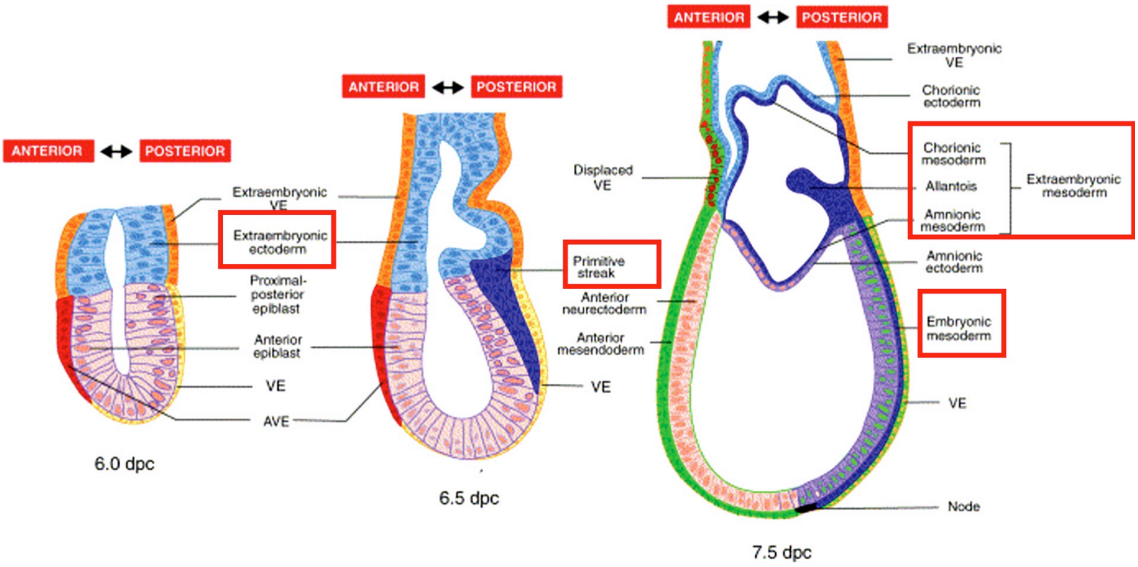

Figure 1.4. Cartoon depicting mouse embryonic development before, during, and after gastrulation commences. Structures highlighted in red boxes are sites where *Bmp4* is endogenously expressed. Adapted from (Lu et al. 2001).

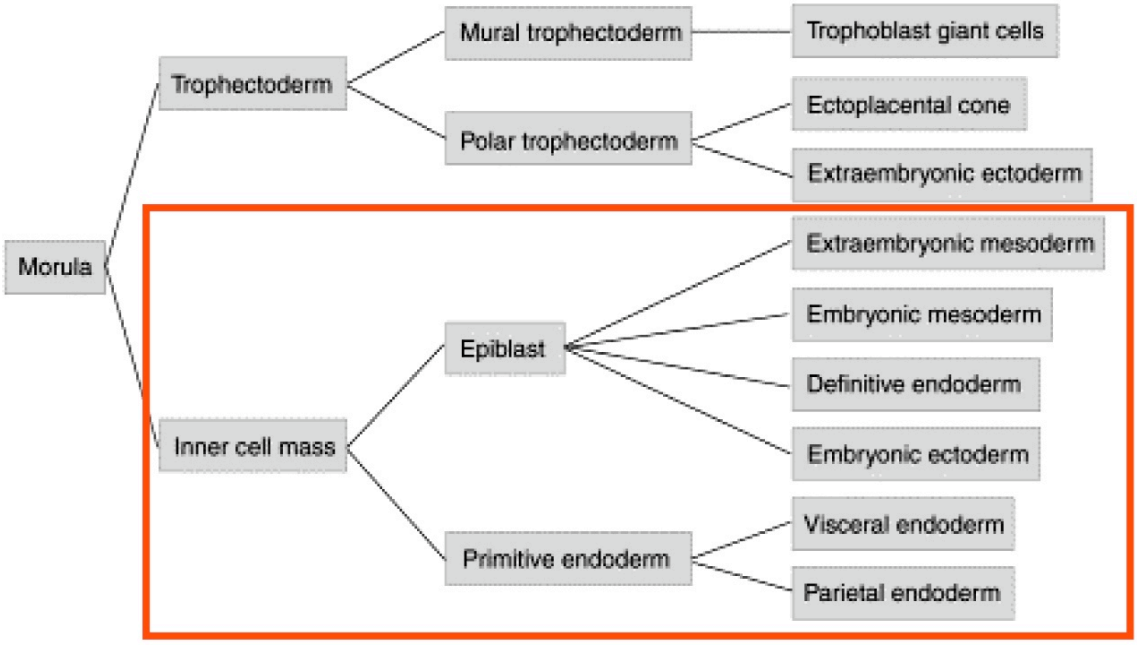

Figure 1.5. Diagram depicting the origins of different cell types in the early mouse embryo. Note, extraembryonic and embryonic mesoderm arise from a common progenitor tissue, the epiblast. Adapted from (Lu et al. 2001).

of the amnion is to provide a fluid-filled environment for embryogenesis to occur (Gilbert 2003). The chorion contributes to the placenta, providing multiple critical functions for embryogenesis such as respiration, nutrition, and immunity (Gilbert 2003). In general, the allantois manages nitrogen waste products. However, the extent to which it does this varies between mammals (Gilbert 2003). The yolk sac generates blood islands providing hematopoietic precursor cells for the developing embryo in addition to supplying the embryo with nutrients for growth and survival (Hogan 1994) (Gilbert 2003). Embryonic mesoderm differentiates to form intermediate, axial, paraxial and lateral plate mesoderm. From these four categories of embryonic mesoderm, numerous tissue types develop including cardiac tissue, cranial tissue, somatic tissue, kidney, gonad, notochord, blood vessels, and gut wall (Hogan 1994) (Tam and Loebel 2007). Together, extraembryonic and embryonic mesoderm tissues contribute significantly to the development of the mouse embryo. In sum, mesoderm arises from the epiblast to form multiple distinct tissues throughout the embryo as well as outside the embryo.

Bmp4 Plays a Critical Role in Mesoderm Development

Not only has expression data implied *Bmp4* regulation is complex, but additional research has indicated Bmp4 plays a critical, dynamic role in mouse development. The mouse knockout revealed *Bmp4* was necessary for mouse development and, as a result, is embryonic lethal (Winnier et al. 1995). Although the phenotype of knockout embryos was variable, the majority of embryos died at

the onset of gastrulation (6.5 dpc) and failed to form mesoderm tissue (Winnier et al. 1995). Embryos that persisted beyond this stage of development exhibited defects in mesoderm development, such as a pronounced decrease in mesodermal cells, or defects in extraembryonic and embryonic mesodermally-derived structures such as blood islands, allantois, ventral-lateral mesoderm, and primordial germ cells (Winnier et al. 1995) (Lawson et al. 1999).

Germline deletion of BMP4 receptors, BMPR1A or BMPR2, further support the importance of BMP4 signaling for mesoderm development because receptor knockouts fail to develop mesoderm or gastrulate (Mishina et al. 1995) (Beppu et al. 2000). In addition, mice lacking downstream intracellular signaling molecules in the Bmp pathway such as *Smad1*, *Smad2*, *Smad4*, or *Smad5* exhibit gastrulation defects or mesodermally-derived tissue defects implying Bmp signaling plays a critical role in gastrulation and mesoderm development (Tremblay et al. 2001) (Lechleider et al. 2001) (Nomura and Li 1998) (Waldrip et al. 1998) (Weinstein et al. 1998) (Chang et al. 1999). For example, *Smad5* null mice are embryonic lethal and display defects in ventral body wall closure (ventral-lateral mesoderm), allantois and primordial germ cells (Chang et al. 1999) (Chang and Matzuk 2001) and *Smad1* null mice exhibit allantois defects and a significant decrease or complete ablation of primordial germ cells (Tremblay et al. 2001). Taken together, this data clearly demonstrates Bmp4 plays a critical role in the formation and development of extraembryonic and embryonic mesoderm.

Bmp4 Plays a Critical Role in Multiple Distinct Tissues Throughout Development

Not only does Bmp4 play an important role in early embryonic development, but it also is important for the normal development of multiple different tissues as development proceeds. Homozygous knockout mice that develop beyond gastrulation have delayed liver bud morphogenesis (Rossi et al. 2001). Analysis of heterozygous knockout mice revealed multiple haploinsufficient phenotypes including skeletal, kidney, seminiferous tubule, urogenital, eye, craniofacial, and pulmonary vascular smooth muscle cell defects (Dunn et al. 1997) (Miyazaki et al. 2003) (Frank et al. 2005).

Specific inactivation of *Bmp4* in the developing heart revealed Bmp4 is required for atrioventricular septation of the heart (Jiao et al. 2003). Likewise, specific inactivation of *Bmp4* in branchial arch mesenchyme and outflow tract myocardium demonstrated the requirement of Bmp4 for outflow tract septation and branchial arch artery remodeling (Liu et al. 2004). In addition to cardiac defects, tissue-specific inactivation of *Bmp4* uncovered the requirement of Bmp4 for digit patterning (Selever et al. 2004) and distal lung epithelium development (Eblaghie et al. 2006).

Studies suggest Bmp4 is capable of promoting multiple biological functions including induction, chemoattraction, apoptosis, proliferation, and differentiation. Induction occurs when two distinct tissue types are juxtaposed to one another allowing paracrine factors from one tissue type to affect the adjacent tissue. For example, Bmp4 signaling that originates from septum transversum mesenchyme induces the adjacent endoderm to begin transcribing liver-specific

genes resulting in liver morphogenesis (Rossi et al. 2001). In addition, the Bmp4 signal can serve as a chemoattractant as demonstrated by its ability to attract ureter mesenchymal cells in an explant culture system (Miyazaki et al. 2003). Studies in the chick eye have suggested Bmp4 promotes cell proliferation in retina cultures and apoptosis in the optic cup to promote eye development (Trousse et al. 2001). Bmp4 has been show to regulate cell differentiation, as well. For instance, *Bmp4* expression in endocrine cells has been shown to block differentiation and maintain cells in a progenitor state allowing the progenitor cell population to increase (Hua et al. 2006). An alternate way Bmp4 regulates cell differentiation is demonstrated by its ability to promote visceral endoderm differentiation by signaling from underlying ectoderm cells in peri-implantation embryos (Coucouvanis and Martin 1999). In regards to Bmp4's function in early mouse development, studies suggest Bmp4 acting from extraembryonic ectoderm is required for epiblast development (Lawson et al. 1999). In this aspect, Bmp4 is believed to serve as an inductive signal for primordial germ cell progenitors and allantois differentiation (Lawson et al. 1999) (Fujiwara et al. 2001). Bmp4 in extraembryonic mesoderm is believed to promote primordial germ cell survival/localization and allantois differentiation (Fujiwara et al. 2001). Overall, Bmp4 exhibits pleiotropic biological functions throughout development.

Taken together, the inability of minimal promoter fragments to recapitulate the complete repertoire of *Bmp4* endogenous expression patterns, coupled with the critical role Bmp4 plays in embryonic development as well as its complex spatiotemporal expression patterns suggests *Bmp4* maintains a complex *cis*-

regulatory architecture allowing for precise control of *Bmp4* expression in multiple different tissues throughout development.

Identification of *Cis*-Regulatory Elements

Studies in model organisms have shed light on the structure of *cis*-acting transcriptional regulatory elements that mediate developmental signals in animals. Numerous studies of both fly and sea urchin *cis*-regulatory sequences indicate that often, individual *cis*-regulatory elements are sequence "modules" of a few hundred base pairs or less in length and are bound *in vivo* by approximately 4 to 8 different types of transcription factors (Davidson 2001). Genes with multiple developmental functions typically have several, separate *cis*-regulatory modules. For example, numerous *cis*-modules of the Drosophila melanogaster (fly) gene, decapentaplegic (dpp), control its expression in different embryonic tissues such as imaginal discs, mesoderm, gut, and brain (Blackman et al. 1991). Mutation analysis and reporter gene constructs were used to localize enhancers in the 5', proximal, and 3' regions of the *dpp* locus (St Johnston et al. 1990) (Blackman et al. 1991). *Cis*-acting regulatory modules may be even more widespread in vertebrate genomic DNA, contributing to organismal complexity. These can be hundreds of kilobases upstream or downstream of the genes they regulate, as suggested by research on *Shh* (Roessler et al. 1997) (Spitz et al. 2003), *Gdf6* (Mortlock et al. 2003), *Bmp2* (Chandler et al. 2007) and *Sox9* genes (Wirth et al. 1996) (Wunderle et al. 1998) making it difficult to localize distant enhancers with conventional techniques. However, *cis*-modules

can often be identified due to evolutionary conservation and have been found within some gene deserts (Nobrega et al. 2003) (Nobrega et al. 2004). The human *Dach* gene is surrounded by two large gene deserts (870kb, 1330kb) and it is expressed in multiple tissues during development (Nobrega et al. 2003). To examine the possibility that functional elements existed in the gene deserts and contributed to *Dach's* dynamic expression patterns, researchers used comparative sequence analysis to extract highly conserved sequences (Nobrega et al. 2003). Seven of nine elements tested for reporter activity in mouse were capable of directing expression patterns that recapitulated a subset of *Dach* endogenous expression patterns (Nobrega et al. 2003). This raises the possibility that other developmentally important genes flanked by gene deserts are regulated similarly. Yet very few gene deserts near developmentally important genes have been studied in detail.

Bmp4 resides in a gene desert of approximately 1-megabase (Mb) (FIGURE 1.6). The noncoding sequence within this desert is peppered with highly conserved regions (CHAPTER II), some of which could be functional enhancers for *Bmp4*. Prior to this study, nothing was known about the function of the desert engulfing *Bmp4*. However, there are clues from research on other Bmps. Evidence suggests that some *Bmp* genes utilize long-range *cis*-regulatory mechanisms to achieve tissue specific gene expression. The employment of transgenic reporter methods in mice allowed researchers to map enhancers residing over 200kb from *Bmp5* (DiLeone et al. 1998). These enhancers were found to control expression of *Bmp5* in specific anatomical locations during

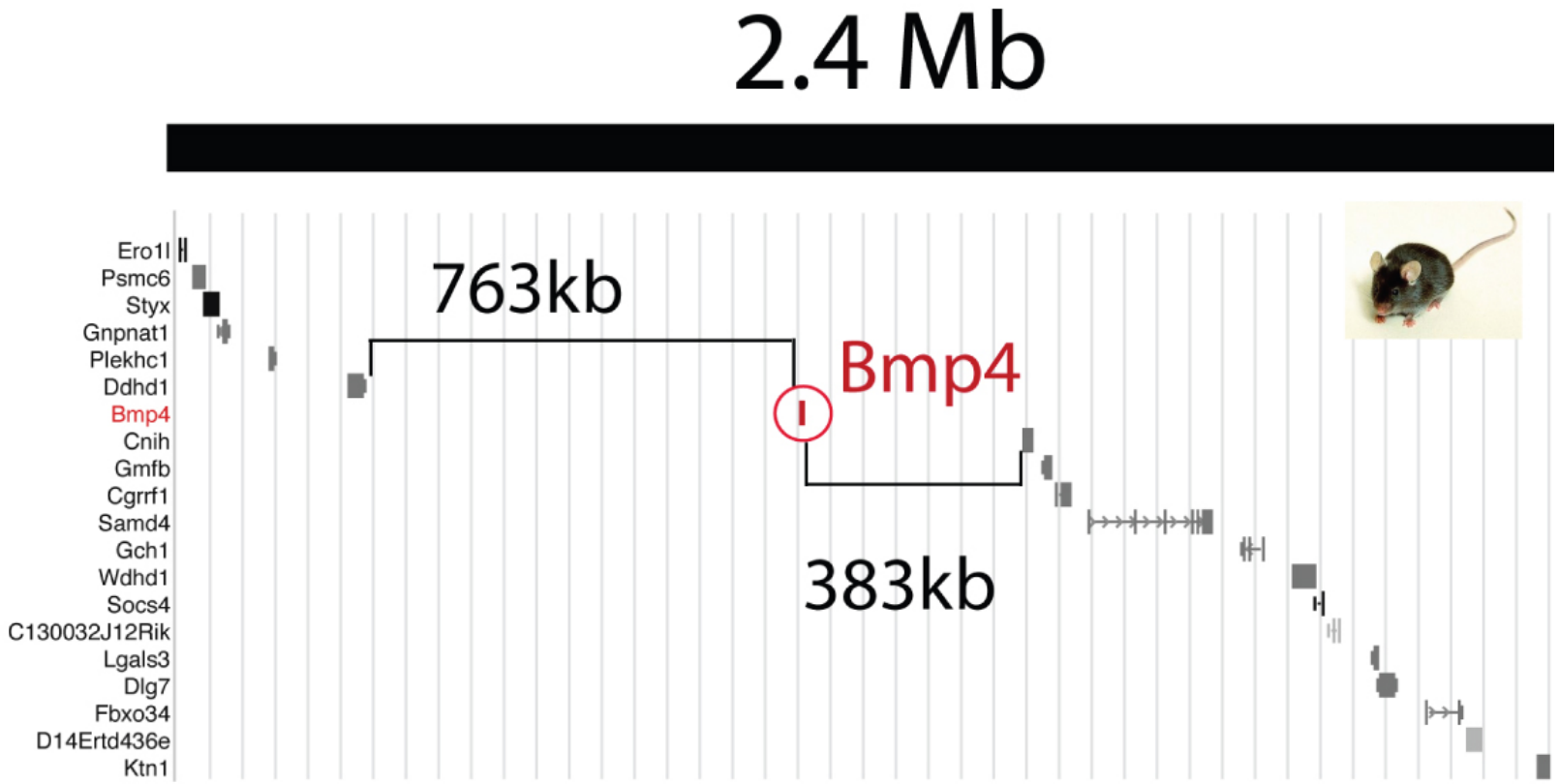

Figure 1.6. *Bmp4* resides in a gene desert. Shown here is a 2.4 Mb segment of mouse chromosome 14. Bmp4 is circled in red and is located on the negative strand (5' is on the right, 3' is on the left). Note, the region immediately 5' (383 kb) and 3' (763 kb) to Bmp4 is devoid of any known protein-coding sequences. Beyond this 1.1 Mb desert, multiple genes are present at regularly-spaced intervals.

embryonic development, such as ribs, sternum and ear cartilage. Likewise, an overlapping bacterial artificial chromosome (BAC) transgenic approach was used to localize distant enhancers of *Gdf6* (*Bmp13*) and *Bmp2* that directed distinct expression patterns (Mortlock et al. 2003) (Chandler et al. 2007).

As with most transcriptional studies, the focus of *Bmp4* regulation has been limited to several kilobases surrounding the proximal promoter. In the case of *Bmp4*, these studies have failed to identify regulatory regions for all expression patterns seen endogenously. In one study, as described above, transgenic reporter mice were created using successively shorter pieces of DNA extending 2.4 kb upstream of the mouse *Bmp4* promoter (Feng et al. 2002) (Zhang et al. 2002). This approach identified a regulatory region that directed only a portion of the endogenous expression patterns (tooth ameloblasts, hair follicles, and somites), indicating that many other sites of *Bmp4* expression are induced by control modules outside the minimal 2.4 kb fragment (FIGURE 1.3) (Feng et al. 2002) (Zhang et al. 2002). Furthermore, stable transgenic fish lines containing a 7.5 kb *Bmp4* promoter/reporter construct failed to recapitulate any endogenous expression patterns (FIGURE 1.3) (Shentu et al. 2003). In addition, work in our lab has shown that the closest homolog of *Bmp4*, *Bmp2*, has multiple enhancers that reside long distances from the gene (Chandler et al. 2007). Therefore, it is likely that long-range regulatory mechanisms play a role in *Bmp4* regulation.

Much evidence suggests that developmentally important *cis*-regulatory elements are often highly conserved across species. This is particularly useful as an indication of candidate regions of DNA that might harbor regulatory

sequences. Evidence also suggests that conserved, noncoding sequences present in gene deserts can represent enhancers for nearby genes (Nobrega et al. 2003) (Nobrega et al. 2004) (Woolfe et al. 2005). Gene deserts that contain conserved, noncoding sequences tend to be near developmentally regulated genes more often than gene deserts that do not have these sequences (Ovcharenko et al. 2005b). In summary, several lines of evidence have suggested that long-range regulatory modules regulate *Bmp4* expression. First, similar findings have been documented for other *Bmp* family genes, including the most closely related *Bmp* family member, *Bmp2*. Second, the regulatory complexity of *Bmp4* has been documented and its importance in development has been demonstrated. Finally, the gene deserts bracketing *Bmp4* contain well-conserved noncoding sequences.

Now that genome sequencing and assembly is completed or nearly completed in multiple different species, comparative analyses can be performed using many different species. For example, genomic comparisons can be made between numerous species such as human, non-human primates, rat, mouse, opossum, dog, chicken, frog, zebrafish, pufferfish and fly. Therefore, analyses can be performed between eukaryotes (eg. human vs. fly), vertebrates (eg. human vs. pufferfish), or placental mammals (eg. human vs. mouse). However, there are advantages and disadvantages to each type of comparison as reviewed by Boffelli and colleagues (Boffelli et al. 2004). Analyses of species that are closely related such as human and non-human primates are advantageous for locating primate-specific cis-regulatory elements. The

disadvantage of human/non-human primate sequence comparison is the extreme sequence similarity between the species resulting in a large number of false-positive results and making it difficult to quickly identify a functional element. In contrast, comparisons between human and a more divergent species, such as mouse, will yield significantly fewer 'evolutionarily conserved regions' or ECRs with a higher likelihood of being functional. Although comparisons between human and mouse yield significantly fewer ECRs than human and non-human primates, this type of comparison will return a large volume of ECRs, many of which are non-functional, most likely due to the variability in evolution rates throughout the human genome. To narrow down the sheer number of ECRs for functional tests, comparisons between more divergent species, such as human and pufferfish, have been fruitful (Aparicio et al. 1995) (Kimura-Yoshida et al. 2004) (Lettice et al. 2003) (Nobrega et al. 2003) (Pennacchio et al. 2006) (Woolfe et al. 2005). Comparisons between two species that are more evolutionarily divergent, such as human and pufferfish, often identify functional *cis*-regulatory elements. However, the disadvantage to this type of comparison is it tends to identify enhancers for structures that are common to both fish and mouse eliminating the potential to identify a mammalian-specific enhancer. Regardless of the potential downfalls some species have in comparative analyses, a combination of sequence comparisons between multiple species will be necessary to identify *cis*-regulatory elements.

Thesis Overview

A significant amount of research on Bmp4 has been published since its discovery nearly twenty years ago, yet little is known regarding the transcriptional regulation of this gene. Expression analysis has suggested that *Bmp4* regulation is complex and functional studies indicate Bmp4 plays a critical role in multiple tissue-specific ways. Analysis of the genomic architecture surrounding *Bmp4* revealed the gene resides in a gene desert, which is often associated with genes that show dynamic spatiotemporal expression patterns requiring numerous *cis*-regulatory elements. In addition, *in vivo* studies have suggested numerous *cis*-regulatory elements reside beyond the minimal promoter fragments that were tested. However, to date, nothing has been published to definitively show that *Bmp4* enhancers are present beyond the minimal promoter. Likewise, the enhancer(s) required for *Bmp4* expression in mesoderm have not been located. Given the necessary role of Bmp4 in mesoderm development, mapping the mesoderm enhancer would significantly contribute to our current knowledge and allow further research towards identification of upstream factors required for *Bmp4* expression in mesoderm.

The focus of this dissertation is to make a meaningful contribution towards the understanding of *Bmp4* regulation as well as to explore the use of comparative analyses and multiple model organisms to quickly pinpoint functional enhancers. Towards this end, Chapter II describes the use of comparative analyses to identify three evolutionarily conserved regions flanking *Bmp4*. In Chapter III, the use of BAC reporter transgenic mice to ascertain *Bmp4*

enhancers is revealed. This Chapter suggests that many, but not all *Bmp4* enhancers reside in a 398 kilobase (kb) segment of mouse Chromosome 14 containing *Bmp4*. Chapter IV describes a reliable method for determining transgene copy number in BAC transgenic lines as well as the importance of performing copy number analysis in BAC transgenic lines as this pertains to transgene integrity and expression. Chapter V details the use of deletion BACs to test for the requirement of evolutionarily conserved regions (Chapter II) for reporter expression. Chapter VI focuses on the use of zebrafish to test each evolutionarily conserved region for enhancer activity, while Chapter VII elaborates on the sufficiency of evolutionarily conserved regions to direct tissue-specific expression in mouse. Finally, Chapter VIII summarizes the research described in this dissertation and makes conclusions based on the research as well as touches on the potential future directions of this project. The data presented in Chapter IV were recently published in the manuscript, "Relevance of BAC transgene copy number in mice: transgene copy number variation across multiple transgenic lines and correlations with transgene integrity and expression" (Chandler et al. 2007b).

Table 1.1. Patterns of endogenous *Bmp4* expression during pre- and postnatal development.

Stage	Expression pattern(s)	Assay	Reference
E3.5	inner cell mass and polar trophectoderm of blastocyst	ISH	Coucouvanis et al. 1999
E4.5	inner cell mass and polar trophectoderm of blastocyst	ISH	Coucouvanis et al. 1999
E5.5	uncavitated extraembryonic ectoderm	LacZ	Lawson et al. 1999
"	uncavitated extraembryonic ectoderm	ISH	Coucouvanis et al. 1999
"	extraembryonic ectoderm	ISH	Ying et al. 2001
E6.0-E6.5 (ES)	extraembryonic ectoderm adjacent to epiblast	LacZ	Lawson et al. 1999
E6.5	posterior primitive streak	ISH	Winnier et al. 1995
"	extraembryonic ectoderm	ISH	Coucouvanis et al. 1999
"	extraembryonic ectoderm	ISH	Ying et al. 2001
E7.25 (MS/LS)	extraembryonic ectoderm within posterior amniotic fold, extraembryonic mesoderm	LacZ	Lawson et al. 1999
LS	extraembrynoic mesoderm	LacZ	Lawson et al. 1999
OB	extraembrynoic mesoderm, allantoic bud, amnion, chorion	LacZ	Lawson et al. 1999
EB	yolk sac mesoderm, chorion, amnion, allantoic bud	LacZ	Lawson et al. 1999
NP - HF	extraembrynoic mesoderm portions of amnion, yolk sac, and chorion, allantois	LacZ	Lawson et al. 1999
E7.5	posterior primitive streak, allantois, amnion, anterior neural region	ISH	Winnier et al. 1995
"	allantois, amnion	ISH	Jones et al. 1991
E8.0	lateral plate mesoderm (6S), *no expression in primitive streak or node*	LacZ	Fujiwara et al. 2002
"	surface ectoderm surrounding neural folds, extraembryonic mesoderm	ISH	Zakin et al. 2004

E8.5	mesoderm/endoderm in posterior primitive streak region, ventral mesoderm, myoepicardium of heart, allantois, amnion	ISH	Jones et al. 1991
"	surface ectoderm and neural folds of forebrain, dorsal midline neuroepithelium of forebrain and midbrain, diencephalon, posterior mesoderm, amnion, mesodermal component of visceral yolk sac, thoracic body wall adjacent to pericardial cavity of heart	ISH	Dudley et al. 1997
"	septum transversum mesenchyme, cardiac mesoderm	LacZ	Rossi et al. 2001
"	heart outflow tract, sinus venosus	LacZ	Jiao et al. 2003
E9.0	neuroepithelium of diencephalon, posterior otic vesicles, dorsal ectoderm of first branchial pouch, anterior portion of space between frontonasal mass and first branchial arch, outer myocardial layer of developing atrioventricular canal of heart, mesoderm surrouding gut and lung bud, somatic and splanchnic mesoderm, mesenchyme of flank adjacent to forelimb bud	ISH	Jones et al. 1991
"	anterior dorsal neuroectoderm	ISH	Furuta et al. 1997
"	distal optic vesicle and overlying surface ectoderm (lens placode). ectoderm of naso-oral region	ISH	Furuta et al. 1998
"	ventral mesenchyme surrounding gut tube (19S)	LacZ	Weaver et al. 1999

"	dorsal midline of common atrium, atrioventricular canal	LacZ	Jiao et al. 2003
"	outflow tract myocardium of heart, myocardium overlying branchial-arch artery junction, aortic sac, mesoderm ventral to branchial-arch arteries, pharyngeal endoderm, branchial arch mesenchyme	LacZ	Liu et al. 2004
"	dorsal telencephalon, eye, prosimal ectoderm of first branchial arch, frontonasal mass, maxillary arch, limb buds, ventral-posterior mesoderm, allantois	ISH	Zakin et al. 2004
E9.5	branchial arches, heart, foregut, posterior ventral mesoderm	ISH	Winnier et al. 1995
"	surface ectoderm overlying doral neural tube, neural crest cells, future neural retina, anterior optic vesicle, presumptive cephalic neural crest cells	ISH	Dudley et al. 1997
"	dorsal forebrain, anterior dorsal roof of telencephalon	ISH	Furuta et al. 1997
"	dorsal tip of optic vesicle, ectoderm of naso-oral region	ISH	Furuta et al. 1998
"	ventral mesenchyme of developing lung (27S)	LacZ	Weaver et al. 1999
"	septum transversum mesenchyme	LacZ	Rossi et al. 2001
"	cardiomyocytes overlying the inferior endocardial cushion, dorsal wall of atrium	LacZ	Jiao et al. 2003
"	ventral limb bud ectoderm, limb bud mesoderm	LacZ	Selever et al. 2004

"	ventral pharynx, third pharyngeal arch core mesenchyme (possibly mesoderm), overlying ectoderm of third pharyngeal arch and cleft, mesenchyme caudal to fourth arch and adjacent to surface ectoderm	LacZ	Patel et al. 2006
"	dorso-distal optic vesicle where neuroepithelium contacts surface ectoderm (presumptive neural retina), mesenchyme ventral to optic vesicle, surface ectoderm	ISH	Behesti et al. 2006
"	mesenchyme of ventral foregut	LacZ	Que et al. 2006
E10.0	oral epithelium	ISH	Aberg et al. 1997
"	ventral mesenchyme of primordial lung buds (30S), endoderm of lung	LacZ	Weaver et al. 1999
"	dorsal telencephalon, eye, prosimal ectoderm of first branchial arch, frontonasal mass, maxillary arch, limb buds, ventral-posterior mesoderm, allantois	ISH	Zakin et al. 2004
E10.5	floorplate of diencephalon, nasal pit ectoderm, distal ectoderm of facial processes, mesenchyme surrounding gut, myocardial layer of truncus arteriosus of heart, apical ectodermal ridge of fore and hindlimbs, mesenchyme of limb bud	ISH	Jones et al. 1991
"	dorsal hindbrain, telencephalon, dorsal midline of anterior diencephalon, optic cup	ISH	Furuta et al. 1997
"	dorsal margin of optic cup, ectoderm of naso-oral region	ISH	Furuta et al. 1998

"	cardiomyocytes overlying the inferior endocardial cushion, muscular layer of atrial septum primum, mesenchyme of truncus cushion, dorsal wall of atrium, venous valves	LacZ	Jiao et al. 2003
"	outflow tract of heart, cardinal veins, coronary sinus, inferior vena cava	LacZ	Liu et al. 2004
"	limb bud mesoderm underlying apical ectodermal ridge (higher expression in posterior vs. anterior), apical ectodermal ridge	LacZ	Selever et al. 2004
"	epithelium at point of fusion between the medial nasal process and the maxillary process	LacZ	Liu et al. 2005
"	third pharyngeal pouch endoderm, posterior portion of fourth pharyngeal pouch, third pharyngeal cleft, neural crest cells adjacent to Bmp4-expressing endoderm	LacZ	Patel et al. 2006
"	dorsal neural retina	ISH	Behesti et al. 2006
E11.0	anterior retina, optic stalk, mesehchyme surrounding eye	ISH	Dudley et al. 1997
E11.5	mesenchyme of facial processes, secretory/ sensory epithelium of developing ear	ISH	Jones et al. 1991
"	mesial epithelium and underlying mesenchyme of palatine rugae (oral surface of maxillary plate), ventral bronchial epithelium, mesenchyme surrounding forestomach/midgut/ bronchi, mesenchyme surrounding urogenital sinus	ISH	Bitgood et al. 1995

"	mesehchyme surrounding stalk of ureteric bud, metanephric mesenchyme	ISH	Dudley et al. 1997
"	forebrain, medial walls of lateral ventricles corresponding to future hippocampus and choroid plexus, roof of diencephalon, anterior dorsal roof between telencephalic hemispheres	ISH	Furuta et al. 1997
"	distal lung endoderm tips, lung mesenchyme	LacZ	Weaver et al. 1999
"	midline mesenchyme of tongue	LacZ	Hall et al. 2002
"	venous valves, muscle layer of outflow tract, muscluar layer of atrial septum primum	LacZ	Jiao et al. 2003
"	limb bud mesoderm underlying apical ectodermal ridge (higher expression in posterior vs. anterior), apical ectodermal ridge	LacZ	Selever et al. 2004
"	ventral and posterior region of bilateral thymus/ parathyroid primordia, mesenchyme adjacent to Bmp4-expression region of endoderm, third pharyngeal cleft ectoderm, third pharyngeal pouch endoderm	LacZ	Patel et al. 2006
"	dorsal-most region of optic cup, mandibular and maxillary processes ventral to optic cup	ISH	Behesti et al. 2006
E12.0	mesenchyme below buccal aspect of dental lamina	ISH	Aberg et al. 1997

"	apical ectodermal ridge, equally in anterior and posterior mesoderm of limb bud, central limb bud mesoderm adjacent to apical ectodermal ridge	LacZ	Selever et al. 2004
"	pancreas	RT-PCR	Dichman et al. 2003
E12.5	mesial epithelia and mesenchyme of tooth germ in maxillary arch, mesial mesenchyme of involuting tooth bud, mesenchyme surrounding epithelium of midgut/hindgut/urethra, myocardium underlying atrioventricular valve and truncus arteriosus	ISH	Bitgood et al. 1995
"	distal mesoderm underlying involuting apical ectodermal ridge, tips of forming digits	ISH	Dunn et al. 1997
"	circumvallate and fungiform papillary placodes of tongue	LacZ	Hall et al. 2002
"	thymus and surrounding mesenchyme	LacZ	Patel et al. 2006
E13.0	dental mesenchyme	ISH	Aberg et al. 1997
"	pancreas	RT-PCR	Dichman et al. 2003
"	pancreatic epithelium	ISH	Goulley et al. 2007
E13.5	mesenchyme of facial processes, whisker follicle primordia	ISH	Jones et al. 1991
"	underlying mesenchyme of whisker placode, mesenchyme surrounding stomach	ISH	Bitgood et al. 1995
"	presumptive glomerular region of developing nephron, future podocyte of glomerulus, Bowman's caspsule layers of glomerulus, mesenchyme lining ureter	ISH	Dudley et al. 1997
"	medial walls of lateral ventricles (fimbria), choroid plexus	ISH	Furuta et al. 1997

"	pancreas	RT-PCR	Jiang et al. 2002
E14.0	dental papilla, dental mesenchyme	ISH	Aberg et al. 1997
"	circumvallate and fungiform papillary placodes of tongue	LacZ	Hall et al. 2002
"	pancreas	RT-PCR	Dichman et al. 2003
E14.5	tooth enamel knot, dental papilla, mesenchyme adjacent to whisker placode and pelage hair placode, mesenchyme underlying palatine rugal epithelium, duodenal mesenchyme, rectal mesenchyme, mesenchyme surrounding bladder and vas deferens, bronchioles	ISH	Bitgood et al. 1995
"	distal tips of lung endoderm	LacZ	Weaver et al. 1999
"	intervertebral annulus fibrosus	ISH	Zakin et al. 2004
E15.0	dental papilla, enamel knot	ISH	Aberg et al. 1997
"	pancreas	RT-PCR	Dichman et al. 2003
"	pancreatic islets	ISH	Goulley et al. 2007
E15.5	pancreas	RT-PCR	Jiang et al. 2002
E16.0	pancreas	RT-PCR	Dichman et al. 2003
E16.5	molar dental papilla, mesial aspects of molar cusps, ameloblasts and odontoblasts of incisor, mesenchyme of incisor, dermal papilla of whisker, precortical cells and inner root sheath of whisker, subjacent mesenchyme of involuted hair follicles	ISH	Bitgood et al. 1995
E17.0	cuspal portion of dental papilla including preodontoblastic layer	ISH	Aberg et al. 1997
"	pancreas	RT-PCR	Dichman et al. 2003
"	pancreatic islets	ISH	Goulley et al. 2007
E17.5	distal lung endoderm tips	LacZ	Weaver et al. 1999
"	pancreas	RT-PCR	Jiang et al. 2002

E19.5	mitral cell layer of olfactory bulbs, olfactory epithelium, olfactory receptor neurons	ISH	Peretto et al. 2002
P0	pancreas	RT-PCR	Dichman et al. 2003
"	pancreatic islets	ISH	Goulley et al. 2007
"	muscle layer of outflow tract, annulus of mitral and tricuspid valves	LacZ	Jiao et al. 2003
"	epididymis, rete testis	ISH	Hu et al. 2004
P1	odontoblasts, differentiating ameloblasts	ISH	Aberg et al. 1997
1 wk	epididymis, rete testis, vas deferens	ISH	Hu et al. 2004
2 wk	epididymis, seminiferous tubules, pachytene spermatocytes	ISH	Hu et al. 2004
3 wk	seminiferous tubules, epididymis	ISH	Hu et al. 2004
5 wk	seminiferous tubules, epididymis	ISH	Hu et al. 2004
10 wk	seminiferous tubules, pachytene spermatocytes	ISH	Hu et al. 2004
Adult	endothelial cells in muscularized and nonmuscularized vessels of lung, vascular smooth muscle cells of lung, airway and alveolar epithelium of lung	LacZ	Frank et al. 2005
"	pancreatic islets	ISH	Goulley et al. 2007

CHAPTER II

COMPARATIVE ANALYSIS REVEALS EVOLUTIONARILY CONSERVED REGIONS FLANKING *BMP4*

Introduction

Identification of functional non-coding sequences amidst genomes encompassing over two billion base pairs (Waterston et al. 2002) is a significant challenge facing the scientific community today. Systematic methods that successfully identify functional non-coding sequences are necessary for annotation of these elements throughout various genomes. Likewise, annotation of functional non-coding sequences is critical for understanding how the genome functions.

Before the human genome was sequenced, most scientists believed it contained more genes than any other genome due to organismal complexity. However, this hypothesis was abandoned when the human genome was sequenced and determined to contain less than 30,000 genes, three-fold less than what was predicted (Lander et al. 2001) (Venter et al. 2001) (Pennisi 2007). The bulk of the human genome is comprised of noncoding sequence. One potential function of noncoding sequence may be to increase the complexity of an organism by contributing to the repertoire of *cis*-regulatory elements. Often, developmentally regulated genes maintain numerous *cis*-regulatory elements (Plessy et al. 2005). In turn, modifications in the *cis*-regulatory elements may induce morphological changes in the organism leading to increased complexity

(Carroll 2001) (Wray 2007). Alternatively, mutations in *cis*-regulatory elements may lead to disease as evidenced by a single base pair mutation in the *Sonic hedgehog* (*Shh*) enhancer which causes preaxial polydactyly (Lettice et al. 2002) (Lettice et al. 2003) (Maas and Fallon 2005). Since conservation of sequence often implies conservation of function, it is important to look beyond the traditional view of gene structure. In support of this, comparisons between human and mouse genomes reveal a significant amount of noncoding sequence is at least 70% conserved over at least 100 bp between species (Loots et al. 2000) (Dermitzakis et al. 2005). In fact, nearly 500 regions of the human genome are “ultraconserved” (100% identity between human/rat/mouse for at least 200 bp) (Bejerano et al. 2004). Thus, noncoding elements may represent an additional functional component of the genome.

Approximately 5% of the mammalian genome represents short stretches (<100 bp) of conserved sequence, not all of which can be attributed to coding sequences, posing a challenge for characterizing noncoding, functional elements (Waterston et al. 2002). Recent studies indicate that comparative sequence analysis between more divergent species, such as mouse and fish, is a powerful way to detect ancient *cis*-regulatory elements (Nobrega et al. 2003) (Ahituv et al. 2004) (Pennacchio et al. 2006) (Woolfe et al. 2005) (Boffelli et al. 2004). *Fugu rubripes*, or pufferfish as they are commonly referred to, are members of the *Tetraodontidae* family of teleost fish believed to have diverged from a common ancestor shared with mammals approximately 450 million years ago (mya) (Ureta-Vidal et al. 2003). The pufferfish genome is highly compact (400 Mb)

compared to the human (3 Gb) or mouse genome (2.6 Gb) and contains very little repetitive sequence, but an equivalent amount of coding sequence (Ureta-Vidal et al. 2003) (Elgar et al. 1996). Therefore, the intergenic regions of the pufferfish genome are concentrated with *cis*-regulatory elements (Muller et al. 2002). The evolutionary distance between pufferfish and mammals coupled with the compact nature of the pufferfish genome makes the pufferfish a powerful model organism for identification of noncoding conserved sequences by comparative analysis.

Recent data highlights a newfound component of the genome designated as a “gene desert”. These are defined as regions of the genome at least 500 kilobases in size that are devoid of known protein coding sequences and they comprise approximately one quarter of the human genome (Venter et al. 2001) (Ovcharenko et al. 2005b). Little is known about the functional significance of gene deserts, which may play an important role in gene regulation. Initial experiments suggest at least some gene deserts impart little to no functional significance as demonstrated by a mouse knockout (Nobrega et al. 2004). Gene deserts are characterized by an increase in LINE-type repetitive elements and a decrease in SINE-type repetitive elements, while the overall density of repetitive elements is not significantly different from other regions in the genome (Ovcharenko et al. 2005b). In addition, gene deserts can be divided into two classes: stable and variable. Stable gene deserts have greater than 2% of their sequence conserved between human and chicken, contain 98% of noncoding sequences conserved between human and fugu, include regions that surround

genes already shown to contain long-range regulatory elements, have a decreased density of repetitive sequence, and maintain synteny with their adjacent genes (Ovcharenko et al. 2005b). Alternatively, variable gene deserts have less than 2% of their sequence conserved between human and chicken and are mammalian-specific (Ovcharenko et al. 2005b). Although these cutoffs are arbitrary, they are useful to describe the different types of gene deserts. Nevertheless, there appears to be a distinction between gene deserts based on structural and/or evolutionary evidence. Thus, gene deserts may serve functional roles. Understanding the functional significance of conserved noncoding sequences and gene deserts may provide insight into the etiology of genetic diseases that fail to manifest mutations in the classical transcription unit.

Towards this end, *Bmp4* resides in a stable gene desert and is regulated in a complex manner making it a suitable candidate for comparative analysis with pufferfish to identify putative *cis*-regulatory sequences. Initial comparative analyses of human versus mouse genomic sequences identified 336 ECRs peppered across the 5' and 3' gene deserts flanking *Bmp4*. To identify potential *Bmp4 cis*-regulatory elements, we utilized the pufferfish genome for comparative analyses. Similar to findings of other groups who identified conserved noncoding sequences by human/pufferfish sequence alignments (Nobrega et al. 2003) (Pennacchio et al. 2006) (Woolfe et al. 2005), we identified six evolutionarily conserved regions (ECRs) with at least 70% sequence identity over 100 bp or more present in human, mouse, and pufferfish. Finally, comparative analyses of a portion of the gene desert encompassing *Bmp4* across multiple species

revealed the ancient ECRs have been conserved in a syntenic group across millions of years of evolution.

Material and Methods

UCSC Genome Browser

To perform comparative analyses on the *Bmp4* locus, the genomic segments between the adjacent 5' and 3' gene to *Bmp4* on mouse and human chromosome 14 were obtained from the UCSC Genome Browser (http:// genome.ucsc.edu) May 2004 assembly while genomic sequence for the pufferfish *Bmp4* locus was obtained from the October 2004 assembly (Kent et al. 2002). Genomic sequences corresponding to mouse *Bmp4* BACs RP23-26C16 (227,097 bp) and RP23-145J23 (227,220 bp) were obtained from the UCSC Genome Browser (http://genome.ucsc.edu) May 2004 assembly (Kent et al. 2002) and used for comparative analyses. Pufferfish ECR sequences identified from VISTA analysis (see below) were subjected to BLAT analysis on the UCSC Genome Browser to locate each ECR in the zebrafish genome. The UCSC Genome browser was used to evaluate the synteny between multiple species in the gene desert encompassing *Bmp4*.

PipMaker

To detect conserved sequences present in large genomic sequences from fish, mouse and human regardless of orientation we used the BLAST-based local alignment program, PipMaker (Schwartz et al. 2000). Each genomic sequence

was submitted to RepeatMasker (Smit et al. 1996-2004) prior to PipMaker analysis to obtain the position of repetitive elements throughout the sequences allowing them to be masked during PipMaker analysis.

VISTA

To detect regions of conservation in pufferfish, mouse and human genomic segments in the same relative order and orientation, we used the global alignment program, VISTA (Mayor et al. 2000). Genomic sequences obtained from the UCSC Genome browser (see above) were submitted to VISTA in the same order and orientation. In addition, genomic sequences from the *Bmp4* locus of pufferfish, mouse and human were submitted to the zPicture visualization and alignment tool which uses the same local alignment program as PipMaker (see above) and is part of the rVISTA suite (Loots and Ovcharenko 2004). ECR sequences for pufferfish, mouse and human were obtained from this analysis. Conserved and aligned transcription factor binding motifs were detected in ECR sequences using rVISTA (Loots and Ovcharenko 2004). The ECR browser was used to visualize conserved noncoding sequences at the *Bmp4* locus in multiple species (Ovcharenko et al. 2004).

TRANSFAC®

To find predicted transcription factor binding sites in conserved sequences, the weight matrix-based MATCH™ tool from the TRANSFAC® database of transcription factors was utilized (Kel et al. 2003) (Matys et al. 2006).

Results

Multiple Noncoding Evolutionarily Conserved Regions (ECRs) are Present in the Gene Desert Encompassing Mouse and Human *Bmp4*

Little is known about the *in vivo* regulation of *Bmp4*. Studies testing proximal promoter fragments in mice have indicated that very few *cis*-regulatory elements reside in the proximal promoter (Zhang et al. 2002) (Feng et al. 2002). Comparative analysis has been shown to be an effective way to identify functional *cis*-regulatory elements in distal flanking regions (Nobrega et al. 2003) (Ahituv et al. 2004) (Pennacchio et al. 2006) (Woolfe et al. 2005) (Boffelli et al. 2004). Therefore, we performed comparative analysis on the gene desert encompassing *Bmp4* focusing predominately on the genomic regions covered by the BAC clones that were used in transgenic mouse experiments outlined in Chapter III. PipMaker and mVISTA analysis of a 500 Kb region surrounding *Bmp4* on mouse chromosome 14 and the syntenic region on human chromosome 14 revealed significant noncoding conservation. This is easily visualized using the ECR Browser (http://ecrbrowser.dcode.org/) (FIGURE 2.1) (Ovcharenko et al. 2004). For this, the segment of mouse chromosome 14 covered by the 5' and 3' BAC transgenes (Chapter III) was the base sequence and it was compared to the syntenic region of the human chromosome. Here, an ECR is defined as noncoding sequence with at least 70% identity along a minimum of 100 bp. Blue peaks represent coding ECRs, yellow peaks represent UTR ECRs, tan peaks represent noncoding intronic ECRs, and red peaks

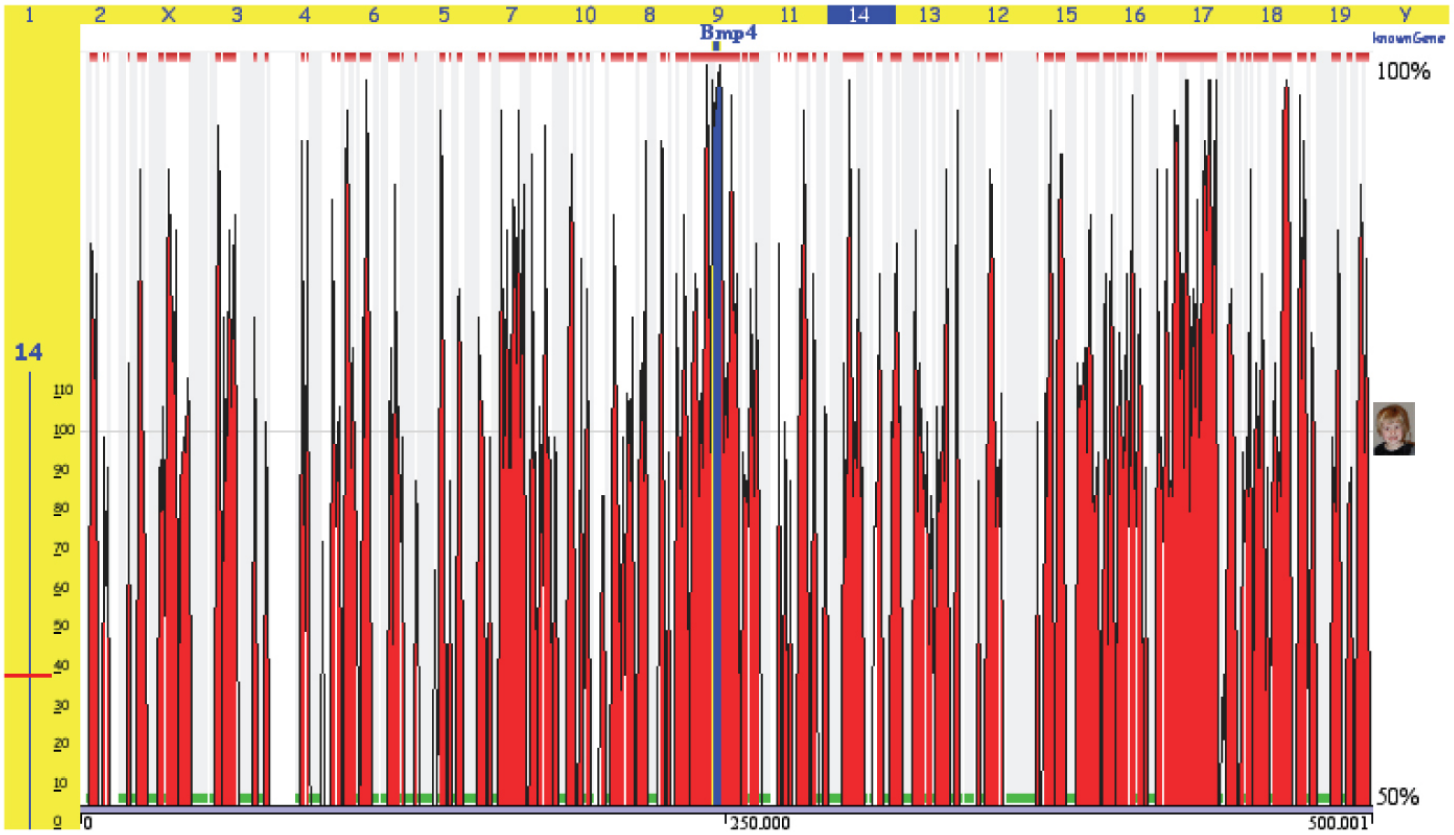

Figure 2.1. Mammalian sequence comparisons revealed hundreds of conserved, noncoding ECRs. Shown here is a 500 kb segment of mouse Chromosome 14 containing *Bmp4* on the ECR browser (Ovcharenko et al. 2004) being compared to the syntenic region in human. *Bmp4* is annotated as a blue peak and noncoding ECRs are denoted by red peaks. Note, the bottom bar represents 50% sequence identity, while the top represents 100% sequence identity over a minimum of 100 bp.

represent noncoding intergenic ECRs (FIGURE 2.1). Over 300 noncoding ECRs are present in this genomic segment (FIGURE 2.1). Although this analysis suggests that many *cis*-regulatory elements for *Bmp4* are spread across this portion of the gene desert, it is difficult to prioritize which elements should be tested for enhancer activity.

Ancient Noncoding Sequences are Present in the *Bmp4* Gene Desert

To identify noncoding ECRs that are most likely to be functional, we utilized comparisons with the pufferfish genome. Others have shown comparisons between highly divergent species, such as pufferfish and human, are most likely to identify functional elements (Pennacchio et al. 2006) (Nobrega et al. 2003) (Woolfe et al. 2005). Although increasing the stringency of the ECR parameters, such as requiring 100% sequence identity over at least 200 bp between mouse/rat/human, has also been shown to identify functional elements (Bejerano et al. 2004), we hypothesized an ancient element conserved over 450 million years of evolution may reveal fundamentally critical elements that are required for normal development since their sequence has been nearly maintained in more divergent (mammal vs. non-mammal) species. For example, ECRs identified by intramammalian sequence comparisons and stringent/ ultraconserved parameters (Bejerano et al. 2004) may identify an important mammalian specific element, such as a hair enhancer. An ancient ECR identified by pufferfish/human (non-mammal/mammal) sequence comparisons are likely to identify an enhancer that both fish and humans require for normal development.

In this regard, ancient ECRs most likely play a role in early developmental processes where there are fewer differences between species. This was the case with an ancient *Shh* ECR that was required for expression of *Shh* in the zone of polarizing activity (ZPA) of the developing limb (Lettice et al. 2003). Although fish have five distinct fin types (dorsal, caudal, anal, pectoral, pelvic) and humans have two distinct limb types (forelimb, hindlimb), the early developmental process of fin outgrowth in the pectoral and pelvic fin bud are essentially the same as limb outgrowth in the fore and hindlimb bud (Mercader 2007). Likewise, both species maintain a ZPA that is critical for normal limb development (Mercader 2007) (Lettice et al. 2003). Thus, comparative analysis between pufferfish and human identified a *cis*-regulatory element specific for a structure that was homologous in fish and human. Identification of this type of element also allows scientists to use both zebrafish and mouse as model systems for future studies.

To identify sequences conserved in divergent species, we used the entire gene desert surrounding *Bmp4* for comparative anlyses. To do this, we obtained genomic sequence from the next adjacent 5' gene to *Bmp4* to the next adjacent 3' gene in each species. Using the BLAST-based local alignment program, PipMaker, we identified six noncoding ECRs present in pufferfish, mouse and human (TABLE 2.1.) Analysis of the same genomic region using the global alignment program, mVISTA, identified the same six ECRs. Three of the six noncoding ECRs identified are located within the 5' or 3' BAC intervals used in our transgenic analysis (Chapter III) (FIGURE 2.2), whereas two noncoding

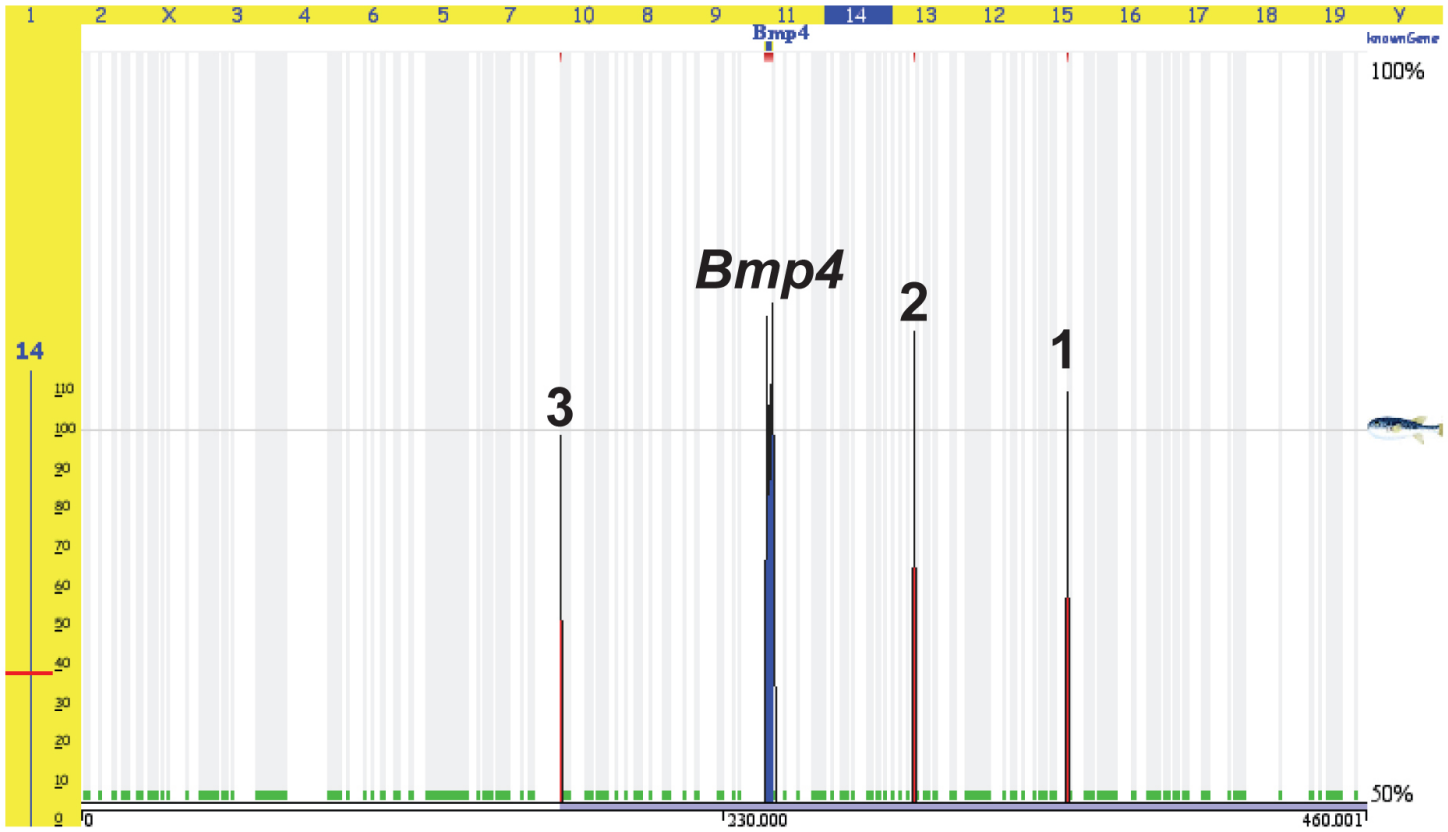

Figure 2.2. Pufferfish/mouse sequence comparisons revealed three conserved, noncoding ECRs. Shown here is a 460 kb segment of mouse Chromosome 14 containing *Bmp4* on the ECR browser (Ovcharenko et al. 2004) being compared to the syntenic region in pufferfish. *Bmp4* is annotated as a blue peak and the three noncoding ECRs present in the BAC intervals are denoted by red peaks. Note, the bottom bar represents 50% sequence identity, while the top represents 100% sequence identity over a minimum of 100 bp.

ECRs are located beyond the BAC intervals tested (TABLE 2.1). For this project, we focused on the three noncoding ECRs present in the BAC intervals. Each ECR has been conserved across multiple vertebrates, including pufferfish, as depicted in an ECR browser plot (FIGURE 2.2). ECR1 and ECR2 are located over 105 Kb and 50 Kb 5' to *Bmp4*, while ECR3 is approximately 74 Kb 3' to *Bmp4* as shown in a zPicture generated from the VISTA suite of comparative tools (FIGURE 2.3). The percent identity between the mouse and pufferfish ECR sequences ranged from 75-81%, with ECR2 being the most highly conserved of the three (FIGURE 2.3). The ancient noncoding sequences are approximately 100 bp in length in their conservation (FIGURE 2.3).

Comparative Analyses Suggest Noncoding ECRs are *Cis*-Regulatory Elements

To look at each ECR sequence in more detail and to evaluate the extent of conservation across multiple species, we took each mouse ECR sequence and performed a BLAT analysis against the mouse genome on the UCSC Genome Browser (FIGURE 2.4). The three grey bars at the top of each BLAT search result indicate that all three reading frames in each ECR contain stop codons (red bars) suggesting these sequences are not translated (note, the opposite strand had the same results) (FIGURE 2.4). ECR sequences may also represent other types of *cis*-regulatory elements, such as locus control regions (LCRs), microRNAs or insulators. According to the UCSC Genome Browser, there are no predicted microRNAs located in any of the three ECR sequences (FIGURE 2.4). CTCF, a zinc finger protein, is required for vertebrate insulator function (Bell

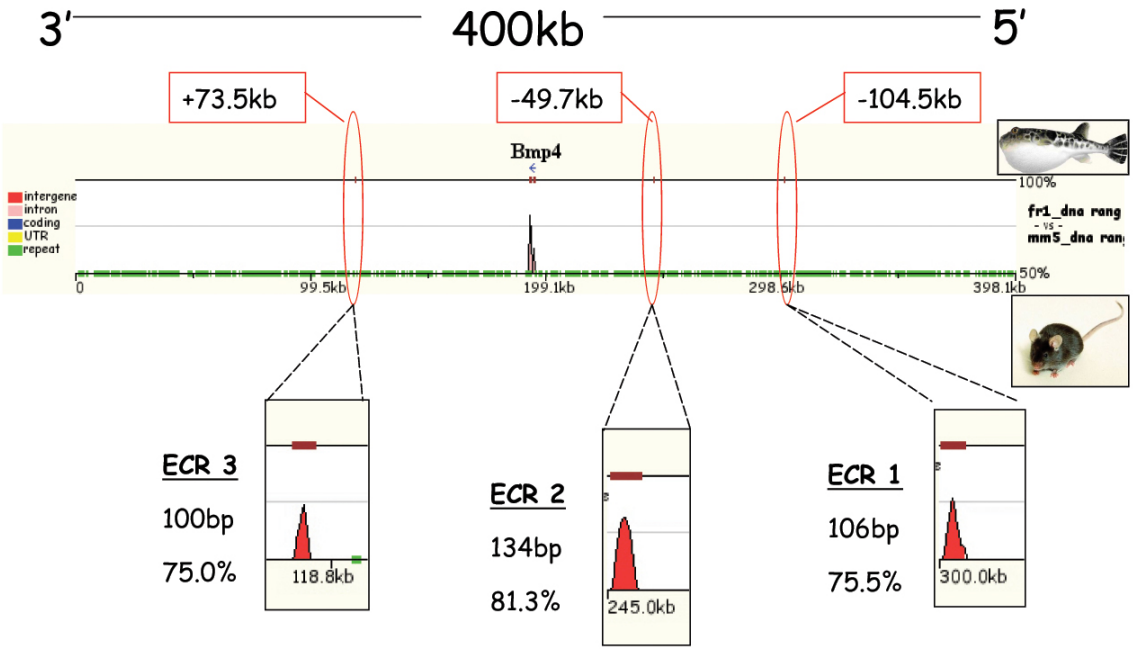

Figure 2.3. Three ancient, long-range ECRs flank *Bmp4*. A graphical representation of ECRs identified by pufferfish/mouse sequence comparisons is depicted using the zPicture portion of rVISTA (Loots and Ovcharenko 2004). Shown here is a 398 kb segment of mouse Chromosome 14 (corresponding to the BAC interval tested in Chapter III) being compared to the syntenic region in pufferfish. Note, *Bmp4* is on the minus strand. ECR1 and 2 reside 104.5 kb and 49.7 kb 5' to mouse *Bmp4*, respectively, while ECR3 resides 73.5 kb 3' to *Bmp4*. Each ECR is at least 100 bp in length and has at least 70% sequence identity between mouse and pufferfish.

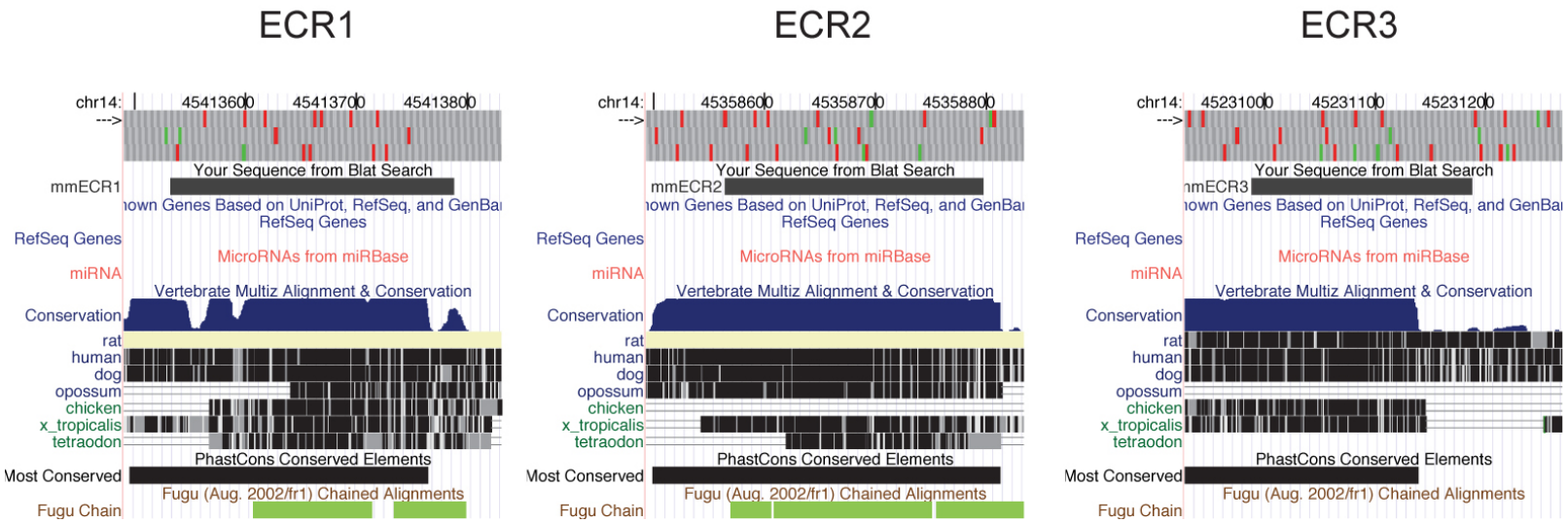

Figure 2.4. Graphical view of mouse ECRs on the UCSC Genome Browser. Each mouse ECR was subjected to a BLAT (Kent 2002) search against the mouse genome on the UCSC Genome Browser (Kent et al. 2002).

et al. 1999). Therefore, we also used two different CTCF consensus sequences (Szabo et al. 2000) in rVISTA analysis to look for conserved or aligned CTCF binding sites in each ECR (Szabo et al. 2000). rVISTA uses transcription factor bind site predictions, sequence comparisons and cluster analysis to locate ECRs and predict conserved and/or aligned binding motifs within ECRs (Loots and Ovcharenko 2004). The first sequence (cccgcynggngg) (Szabo et al. 2000) did not align anywhere in the BAC interval. The second CTCF consensus sequence (ccctc) (Szabo et al. 2000) aligned in multiple locations throughout the 400 Kb BAC intervals. However, there were no CTCF consensus sites present in any of the ECRs. Interestingly, rVISTA analysis revealed conserved transcription factor binding sites in all three ECRs consistent with the idea that they may function as long-range enhancers. ECR1 contained conserved transcription factor binding motifs for Activating transcription factor 3 (Atf3) and Forkhead related activator 2 (Foxf2) sites; ECR2 contained Cebp-delta, Nkx6-1, Engrailed (En1), and Muscle segment homeobox protein-1 (Msx-1) binding motifs; and ECR3 contained Gata1-3, Lmo2, Ppar-gamma, and X box binding protein-1 (Xbp-1) binding motifs. Finally, each ECR sequence has additional flanking sequence that has been conserved across multiple vertebrates as shown by the vertebrate Multiz alignment and conservation track on the UCSC Genome Browser (FIGURE 2.4). Therefore, comparisons between mouse and pufferfish genomic sequences resulted in the identification of three ancient noncoding ECRs that are likely to function as *cis*-regulatory elements.

Comparative Analyses Suggest Syntenic Conservation of ECRs Across Multiple Species

Since our analyses have indicated that three noncoding sequences located in the BAC interval are conserved across multiple species, we wanted to determine if the ECRs were conserved as a syntenic group across species. Therefore, we used the UCSC Genome Browser to look at the order and orientation of each ECR in relation to *Bmp4* in human, mouse and pufferfish as well as the next adjacent genes to *Bmp4*. We found that each ECR is located in the same order and orientation relative to *Bmp4* in all three species (FIGURE 2.5). The distance of each ECR from *Bmp4* was similar between human and mouse, but dramatically different in pufferfish (FIGURE 2.5). Each ECR was approximately 7-10 fold closer to *Bmp4* in pufferfish, which is consistent with the highly compact nature of the pufferfish genome. In both human and mouse, the next adjacent genes are *Cdkn3* (5') and *Ddhd* (3'), respectively (FIGURE 2.5). In pufferfish, however, the next adjacent 5' gene is *Ddhd*, not *Cdkn3*, and the next adjacent 3' gene is *Lbh* suggesting there is a synteny break outside of the three ECRs that has occurred across evolution (FIGURE 2.5). Interestingly, all three ECRs have been maintained as a syntenic group across 450 million years of evolution.

Each ECR is Present in the Zebrafish Genome

Prior to testing each ECR for enhancer activity in zebrafish (Chapter VI), we first wanted to verify that each ECR was present in the zebrafish genome. Due to the incomplete nature of the zebrafish genome assembly, each pufferfish

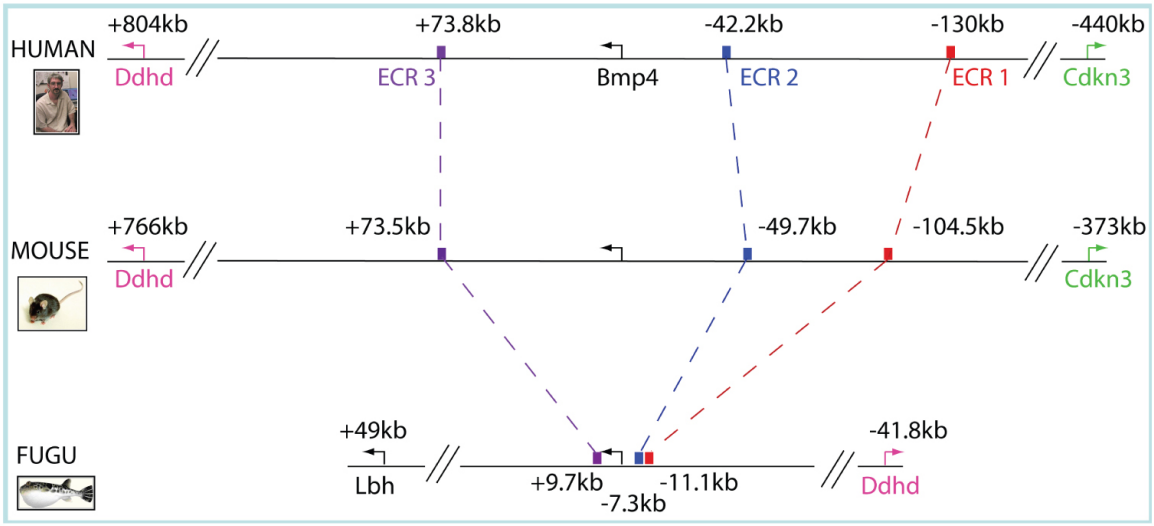

Figure 2.5. Ancient ECRs exhibit syntenic arrangement in human, mouse and pufferfish. The order and orientation of each ECR relative to *Bmp4* suggests ECRs have been maintained in a syntenic block. Interestingly, the adjacent 5' and 3' genes are conserved in human and mouse. However, there is a synteny break upstream of the nearest 3' gene (*Lbh*) in pufferfish. In addition, the adjacent 3' gene (*Ddhd*) in human and mouse is orthologous to the adjacent 5' gene in pufferfish suggesting an inversion occurred outside the syntenic ECR block.

ECR sequence was used to perform a BLAT analysis on the UCSC Genome Browser to locate the ECR sequences in the zebrafish genome. ECR1 was located on contig Zv4_NA6087.1 and both ECR2 and ECR3 were located on contig Zv4_scaffold2294 on the UCSC Genome Browser, whereas *Bmp4* is located on a third contig (Zv4_NA7775.1). In addition to locating each ECR in the zebrafish genome, we performed additional analyses to identify a zebrafish BAC containing the *Bmp4* promoter for potential future studies. We located the contig containing *Bmp4* (Zv4_NA7775.1) on Ensembl (Hubbard et al. 2007) and found the SP6 end sequence of BAC DKEY-255H18 located between two *Bmp4* coding exons. To determine the orientation of the end sequence, we performed a BLAT search alignment and found the end sequence aligned to the minus strand. Therefore, the *Bmp4* promoter should be located on this zebrafish BAC and it was obtained from the Daniokey BAC library for potential future experiments.

Discussion

Multiple groups have shown that comparative analysis of distant species such as mouse and fish is a useful tool for identifying functional *cis*-regulatory elements (Nobrega et al. 2003) (Ahituv et al. 2004) (Pennacchio et al. 2006) (Woolfe et al. 2005) (Boffelli et al. 2004). Genes that are located in gene deserts and display complex expression patterns can be surrounded by numerous *cis*-regulatory elements spread throughout the vast expanse of noncoding sequence (Ovcharenko et al. 2005b) (Sandelin et al. 2004) (Nobrega et al. 2003) (Mortlock et al. 2003) (Plessy et al. 2005) (Chandler et al. 2007). Minimal *Bmp4* promoter

fragments have been tested in mouse, but fail to direct many known patterns of *Bmp4* expression (Feng et al. 2002) (Zhang et al. 2002) suggesting that many *Bmp4* *cis*-regulatory elements reside beyond the minimal promoter. Taken together, we tested the hypothesis that comparative analysis methods would identify noncoding, conserved sequences present in the gene desert encompassing *Bmp4*. Furthermore, we tested the hypothesis that comparisons between mouse and pufferfish would fine tune the detection of *cis*-regulatory elements and allow us to focus on fewer sequences with the most potential for function.

By employing mammal-specific comparative analysis methods, we found hundreds of conserved noncoding sequences in the gene desert surrounding *Bmp4*. These results are similar to other published reports describing the identification of conserved noncoding sequences by comparative analyses across mammalian species (Loots et al. 2000) (Bejerano et al. 2004) (Sandelin et al. 2004) (Dermitzakis et al. 2005) (Nobrega et al. 2003). This is also consistent with the idea that developmentally regulated genes, such as *Bmp4*, are more likely to use multiple long-range *cis*-regulatory elements to maintain their precise and dynamic spatiotemporal expression patterns (Sandelin et al. 2004) (Mortlock et al. 2003) (Plessy et al. 2005) (Chandler et al. 2007b). Although the identification of a large number of potential *cis*-regulatory elements by comparative analysis of mammalian genomes is interesting, comparative analysis of more distant vertebrate genomes, such as pufferfish and mouse, was very effective at locating ancient noncoding sequences. This result has been

replicated with other genes, lending weight to the utility of pufferfish/mammalian comparative analyses (Woolfe et al. 2005) (Nobrega et al. 2003) (Pennacchio et al. 2006). Interestingly, *Bmp2*, the close homolog of *Bmp4* does not seem to have any noncoding conservation present in fish and mammals (Ron Chandler, personal communication). Since these two genes are believed to have arisen from a common ancestral gene, it may suggest that the regulatory landscape surrounding *Bmp4* is more closely related to the fly gene, dpp. Although the high degree of conservation exhibited by each ECR is suggestive that these sequences are maintained across millions of years of evolution due to functional significance, functional tests such as enhancer assays or germline deletions of each ECR are required to prove functionality.

There are multiple types of *cis*-regulatory elements including locus control regions, insulators, promoters and enhancers (Li et al. 1999) (Dermitzakis et al. 2005) (Bondarenko et al. 2003) as well as conserved sequences such as microRNAs or proteins, yet we suggest that the ECRs identified are likely long-distance enhancers due to the presence of multiple transcription factor finding motifs in each sequence, lack of CTCF binding motifs in each sequence that would suggest an insulator function, absence of predicted microRNAs, and presence of multiple stop codons in each alternative reading frame of all ECRs . Given the rVISTA results for each ECR, we can speculate that ECR3 may play a role in directing expression of *Bmp4* in tissues that are part of the hematopoetic pathway (Gata1) (Majewski et al. 2006) (Gata2) (Charles et al. 2006) (Gata3) (Pandolfi et al. 1995) (Lmo2) (Warren et al. 1994) or in (Pparg) placental

development (Imai et al. 2004) given the described function of each transcription factor. Likewise, ECR2 could be hypothesized to direct *Bmp4* expression during neural development (En1) (Wurst et al. 1994). It will be interesting to determine whether hypotheses originating from rVISTA analysis are upheld or disproved by functional assays. Alternatively, we cannot rule out by comparative analysis alone that CTCF binding sites are not present in any of our ECRs. The consensus sequences used for analysis may not be very effective at identifying functional CTCF motifs, or the ECR sequences identified may functionally extend beyond the borders of conservation and contain CTCF binding sites that were in close proximity. Nevertheless, it is imperative that functional assays are performed to ascertain the function of each sequence.

Comparative analyses were instrumental in the identification of the ECR synteny amongst vertebrates. This finding substantiates the hypothesis that *Bmp4* resides in a stable gene desert (Ovcharenko et al. 2005b) and has been shown in *Sonic hedgehog* (*Shh*), another developmentally regulated gene (Goode et al. 2005). The syntenic break at the adjacent genes in pufferfish further supports the conservation of ECR synteny and lends weight to their suggested functional significance.

Table 2.1. Six noncoding ECRs are present in the gene deserts flanking *Bmp4* in pufferfish, mouse, and human

	Mouse/ Fugu % Identity	Distance from *mBmp4* Promoter	Mouse Sequence	Fugu Sequence
ECR1	76%	-104.5 Kb	GCCATTAATGGGCCACATCATCCTCCAC TCTGAAGCAACAGAAGCGGCCAGCTGCT GGCCAGCGAATGAGCGCTGTCTCCAGTG TAAACGTGGCTGATATCATCGCCATCAG AAGGTCATAAAACTGATTGATTATAAAC ACAGCTCTCATTTGTTTACAGTTGACAT TTTTGGGTTTGGGGTAATTAGCTCATTA GCCATCCGTCTCTGGGGGTGAGGGCTCA CAGGTGCTGGCAAGAGCCTTGCAAGGGA AACC	TTTCCGCCATAAATTTGCACTGA TATCACCTCCATCAGAGGGCCAT AAAGCTGATTTATTATAAATACA GCTCTCGTGTGTTTACCATGCAC ATTCTTGGGCCTCCG
ECR2	81%	-49.7 Kb	AACTGTGTCTCTTCAAAACTGACATTTA ATACAGGGTGTAAATCACTAGGGGGGCC TGTTACTCCTTCAGGAATGCAATTACAT AGGGTCAAATAAAACATGAAAGAAACCA ACAAGTTTTAAAATGTAGTATGTGCTCA AAAGCAGGCAGGAAGGAATTCGAAGCAG CCCAAAAATGCTAACACTCACTGCTGTC CTGCGAGAGGGCTGGGAGAAGAGGGCTC TTCACGGTAA	GGGGCTTGTTATTCCTTCAAGAA TGCAATTACATAGTTTCAAATAA AACTTGAAAGAGCACAACAAGTT TTAAAATGCAGTGTGCTCTCCCA AGACAGCAGGCAGGAATTCCAAG CAGTTAGAATATGCTAACA
ECR3	75%	+73.5 Kb	AAGCCCCGGGCCACTTACAATAAAATAA GAGGGAAGCCCAAAGAACCGAAGAAACA AAGGAGATGATTAAGAGATAAGGATCAG TCTCAGATATGACATTTCATCCCGATCA GATGCTAAGCAGATGCTCATAAGACGAA ACACAAAGACAGTTGCTCACTTGACCTT ATCTTTGTGTTCCTCCTTTTTTTCTTTC TCTT	GAGACAAAGGAAGTGATTAAGAG ATAAGGATCAGACTAAGATATGA CATTTCATCCTGGGCAGATCCGG TAGACACACTCCATTCAGACAAT ATATAGCAAC
ECR4	74%	-215.3 Kb	ATAAATGATTAGGGTCATTCTCTAAGAC AAATTTACTTTTCTGCACACTCTAAATC TGTTTAGTCAGTTTAAAGGTGTCAATTA TTGGAACAGGCTAAGACTATGAATTTCT TAGTAATAGCTTTAATAACGTTTTGAAA TGAAAAGCAGATACCTTGAAATCAGCCT GCTTTGAAAAGACCTATGAGGGGCTCTG GTTTTCATTTGTGTGGCGTTCTACATCT GGAATCCCTTATAATTAAAACTAAAAAC CGAACTACTACTTCCACCTGTCAAATGA TGAAAAGATGGATGCACTAATAAGTTAC T	ATAGATGATTAGGGTCATTCTGT GAGACAAATGTAACTTTTCTGGA TGCTCTAAATCTGTTTAGGCACA TTTAATAGTGTCAATTATTGTGG GGCCAGATGGGGGACTCGGGCTT TTCTGTATTCAGGGCCCTTAATA ACATCGGCTACAATGAGAAGCAG ATGCTGACCCGCTTGGTCCCTTT GAAAAAAAGACATTTGAGTAGCT CTGCTTCTAATTTATGAGCCCTC CTTCACATCTGGAATTCTTTATA ATTAAAACATAAACCCCAAAGTT CTACTTCCACCTGTCAAATGTTG AGAAGATGAACATCATAATGGGC TTCT

ECR5	70%	-214.7 Kb	GGCGCAAAGCAATCCCGGACTTCACTGG GGCTGCAGGGAGGGGGGAGGGGAAGGGC ATTATGAACCATGAAGCTGCCAAATTAA CTTCTCACCCTTTCAACCACACTCCAGA GAAT	GGTGAGCAGGGATAGTGGAGTGT TTTGAGGTTAGAGGGTTTGGAGG AGCGAGGTGGCTGAGGGGAAGCA GAGAGTGTTTAGTCCTGAGAGGA AACATTAGCAAAGCGAGGAGGAA AAGCTGCCAAATTAACTTCTCAC CCTTTCAATCAC ACCCATGAGGAT
ECR6	71%	+370.7 Kb	AAGACTTTGTTTTAAATCATAATAGCAA CCAGGCTGTAGTCACATTAGAGCGCACA TGGAGAACTTGCACGAAGTCATGTGTTT TGGATTTGACATTCTGATTTATGGTT	AAGCTGTTGTTTTAAATCATAAT AGTAACCAGACTATAATCACATT ATTGCGCACGTGGTGATTTTGAC AAGTGGTGGGAGTCATGTGCTGA GGCTTTGGCAGGGT

CHAPTER III

BMP4 LACZ-BAC REPORTER TRANSGENES ARE SUFFICIENT TO DIRECT MULTIPLE SITES OF *BMP4* EXPRESSION IN TRANSGENIC MOUSE LINES

Introduction

Bone morphogenetic protein 4 (Bmp4) is a member of the transforming growth factor-beta (Tgfβ) superfamily of secreted signaling molecules. *Bmp4* and its homolog, *Bmp2*, are believed to have arisen from a common ancestral gene (Wozney et al. 1988). Given the high amino acid identity in the mature region of human *Bmp4* and *Bmp2* (92%) and *Bmp4* and *dpp* (76%) (Kingsley 1994), an obvious question to address would be if the proteins function interchangeably. In fact, dpp is sufficient to induce bone in the rat subcutaneous bone induction model as previously demonstrated with Bmp4 and Bmp2 (Sampath et al. 1993). Likewise, human *Bmp4* sequence of the mature coding region in place of *dpp* mature coding sequence is sufficient to rescue dorsal-ventral patterning defects exhibited by *dpp* null embryos (Padgett et al. 1993). Therefore, despite 990 million years of evolution (Ureta-Vidal et al. 2003) the mature coding region of *Bmp4* and *dpp* has been maintained at both the sequence and functional levels.

Due to the evolutionary history of *dpp* and *Bmp4* and ability of *dpp* and *Bmp4* to function interchangeably, their transcriptional regulation may also share similarities. Prior to the cloning of *dpp*, genetic experiments in Drosophila melanogaster first suggested the *dpp* locus is spread over 100 kb of genomic

DNA (Spencer et al. 1982). Mutations in the *dpp* locus display a range of phenotypes from mild imaginal disc perturbations to embryonic lethality suggesting *dpp* expression is critical for multiple distinct developmental processes in Drosophila (Spencer et al. 1982). These genetic studies allowed scientists to begin mapping critical regions of the *dpp* locus. Spencer and colleagues hypothesized that some mutations may represent coding mutations while others may be noncoding mutations in critical *cis*-regulatory regions (Spencer et al. 1982). Subsequent analysis of the *dpp* locus revealed putative *cis*-regulatory elements distributed throughout the locus with some elements residing greater than 30 kb from the promoter (St Johnston et al. 1990) (Masucci et al. 1990) (Blackman et al. 1991). Mutations in specific *cis*-regulatory elements (Blackman et al. 1987) (St Johnston et al. 1990) resulted in altered levels of *dpp* expression in a tissue-specific manner (Masucci et al. 1990) (Masucci and Hoffmann 1993). Furthermore, reporter constructs tested *in vivo* demonstrated multiple noncoding sequences function as tissue-specific enhancers and a subset of these can act at great distances from the promoter (Blackman et al. 1991) (Masucci and Hoffmann 1993) (Huang et al. 1993) (Jackson and Hoffmann 1994). Thus, analysis of *dpp*, the fly homolog of *Bmp4*, revealed the gene maintains a complex array of *cis*-regulatory modules that impart precise spatiotemporal expression of *dpp* throughout fly development. Since *dpp* is the fly homolog of *Bmp4*, *Bmp4* may also require numerous *cis*-regulatory elements that regulate its specific spatiotemporal expression throughout mouse development.

Bmp4 is known to be involved in multiple developmental processes including dorsoventral patterning, gastrulation, and organogenesis (Kingsley 1994) (Hogan 1996). In addition, *Bmp4* displays precise spatiotemporal expression patterns throughout development (CHAPTER I). Therefore, *Bmp4* transcription is likely to be complex since the gene is specifically transcribed in discrete patterns and times during mouse development. One mechanism that enables a gene to maintain a complex transcriptional profile is the employment of numerous *cis*-regulatory elements distributed throughout the gene locus. *Cis*-regulatory domains are noncoding DNA sequences, typically being approximately 100-1000 bp in size with 6-15 transcription factor binding sites that bind 4-8 different transcription factors, resulting in the activation or repression of transcription of a gene in *cis* (Davidson 2001) (Wray et al. 2003). Understanding a gene's *cis*-regulatory architecture can provide insight into upstream factors that regulate gene expression as well as the impact evolution has had on the *cis*-regulatory landscape of a gene. Despite the importance of mapping the regulatory landscape of *Bmp4*, little is known about its regulatory landscape. Experiments in mouse indicate few *cis*-regulatory elements reside near the *Bmp4* promoter suggesting the *cis*-regulatory landscape of *Bmp4* is widespread (Zhang et al. 2002) (Feng et al. 2002). This hypothesis is strengthened by 1) the dynamic developmental expression patterns displayed by *Bmp4* as well as the critical role *Bmp4* plays in many developmental processes (CHAPTER I), 2) increasing evidence that developmentally regulated genes can maintain complex, widespread *cis*-regulatory landscapes (Sandelin et al. 2004) (Plessy et al. 2005)

(Gomez-Skarmeta et al. 2006) (Chandler et al. 2007) (DiLeone et al. 1998) (Wunderle et al. 1998) (Kimura-Yoshida et al. 2004) (Lettice et al. 2003) (Lettice et al. 2002) (Mortlock et al. 2003) (Nobrega et al. 2003), 3) the fly homolog of *Bmp4* (*dpp*) utilizes *cis*-regulatory elements dispersed throughout its distant 3' landscape (see above), 4) the presence of numerous noncoding conserved sequences in *Bmp4's* genomic landscape (CHAPTER II), and that 5) *Bmp4* resides in a gene desert similar to other genes that maintain complex *cis*-regulatory architecture (CHAPTERII) (Ovcharenko et al. 2005b).

To further explore the *cis*-regulatory landscape of *Bmp4*, we assayed the transcriptional activity of large, partially overlapping segments of DNA containing *Bmp4* in mice using BAC reporter transgenes. We hypothesized *Bmp4* maintains numerous, separable *cis*-regulatory modules dispersed throughout a large genomic region. Although others have demonstrated the sufficiency of the proximal *Bmp4* promoter's ability to direct some sites of *Bmp4* expression *in vivo* such as distal whisker matrix, hair shaft and tooth ameloblast of newborn mice (Zhang et al. 2002) (Feng et al. 2002), we focused our efforts on analyzing the sufficiency of *Bmp4* BAC reporter transgenes to direct lacZ expression in sites where *Bmp4* is known to be endogenously expressed during prenatal mouse development. By utilizing two *Bmp4* BAC reporter transgenes that extend as far 5' or 3' as possible while still containing the *Bmp4* transcription unit, we show *Bmp4* maintains a complex *cis*-regulatory landscape with some enhancers located within 25 kb of the promoter and numerous other enhancers located over 25 kb 5' or 3' to *Bmp4*. Interestingly, our results also suggest that some

enhancers reside beyond the confines of the 400 kb segment covered by the *Bmp4* BAC reporter transgenes indicating *Bmp4* may utilize *cis*-regulatory elements that are located over 200 kb from the promoter.

Material and Methods

BAC Reporter Transgenes

We used the UCSC Genome Browser (Kent et al. 2002) to identify two mouse bacterial artificial chromosome (BAC) clones that extend as far 5' and 3' relative to *Bmp4* as possible while still containing the transcription unit. Mouse *Bmp4* BACs RP23-26C16 (227,097 bp) and RP23-145J23 (227,220 bp) were obtained from Children's Hospital Oakland Research Institute (CHORI) (http://bacpac.chori.org/) and verified using restriction enzyme digestion with a frequently cutting restriction enzyme (BamHI) followed by gel electrophoresis on a 0.8% agarose Tris-Acetate-EDTA (TAE) gel overnight at 50 V. The resulting fingerprint gel was compared to the expected banding patterns of each insert sequence. Banding pattern predictions were done using MacVector. Additional verification was done using rare cutting restriction enzymes (NotI, SalI) followed by pulsed field gel electrophoresis (PFGE) on a 1% Tris-borate-EDTA (TBE) gel (6 V/cm, switch time=0.2-22 seconds, 18°C, pump speed=60, run time= 15 hours) to verify the expected banding pattern.

Briefly, the wild-type BAC DNAs were first purified using Clontech Nucleobond BAC maxiprep kits (Catalog #635941) and quantified by analysis of SalI restriction digested DNAs on a pulsed-field gel as compared to HindIII-digested

lambda DNA standards of known mass which were loaded in the adjacent lanes. The pulsed-field gel was run as described above. The gel was stained with ethidium bromide, destained for at least 30 minutes in water, imaged using a BioRad GelDoc, and BioRad GelDoc software was used to determine the DNA band intensities and the estimated masses of DNA in BAC bands relative to the standards. To calculate the mass of total BAC DNA loaded per lane, the estimated mass of the SalI band was multiplied by the ratio of the predicted size of the total BAC (see above) to the predicted size of a BAC restriction digest band. SalI digestion of either BAC produces a 6.4 kb band liberated from the BAC vector backbone, which was useful for this calculation. The total mass of BAC DNA per lane was divided by the volume of BAC DNA loaded per lane resulting in the concentration of purified BAC DNA.

The BAC DNAs were then transferred into EL250 cells. Electrocompetent EL250 cells were prepared as follows: On day 1 EL250 cells were streaked from a glycerol stock onto an LB (Luria-Bertani broth) plate with 25 μg/ml chloramphenicol (CAM) and incubated overnight at 32°C. On day 2, a single colony was innoculated into 2 ml of liquid LB+CAM and incubated overnight at 32°C with shaking. On day 3, 0.4 ml of the miniculture was used to inoculate 20 ml LB+CAM. The culture was incubated with shaking at 32°C until the OD (optical density) of 1 ml samples reached an A600 (absorbance at wavelength 600 nanometers) value of 0.4 as measured with a spectrophotometer (Amersham GeneQuant Pro). Next, the cells were incubated on ice for 20 min. For all following steps the cells were kept on ice and tubes and 10% sterile glycerol were pre-

chilled on ice; all centrifuge steps were are at 4°C. The cultures were then transferred to prechilled 50 ml tubes and the cells pelleted by centrifuge at 5000 rpm for 10 min. The supernatant was discarded, cells resuspended in a total volume of 3 ml with prechilled sterile Millipore-filtered deionized water and split into two microcentrifuge tubes on ice. The cells were pelleted in a microcentrifuge at 5000 rpm for 5 minutes, and the pellet was rinsed three times more with 1.5 ml prechilled sterile water. The final pellet was resuspended in ~35 µl of sterile-filtered 10% glycerol/90% deionized water, snap-frozen on dry ice and stored at -80°C until use. To electroprate the BAC DNAs, 200 ng of each BAC in 1.0-2.0 µl of TE (10 mM Tris-HCL [pH 7.4], 1 mM EDTA [pH 8.0]) were used for electroporation into one aliquot of EL250 cells using a BioRad Gene Pulser Xcell and Gene Pulser/MicroPulser Cuvettes (0.1 cm gap) (Bio-Rad Life Science, catalog #165-2089) using 1.8 kV/200 ohms/25 microfarads/capacitance extender set to “off”. After pulsing , 960 µl of LB were added and the cells recovered with shaking for 90 minutes at 32°C . The cells were then plated on LB+CAM plates and incubated overnight at 32°C . Isolated colonies were picked and BAC DNA preps were made and analyzed by restriction digest to confirm transfer of intact BACs into the EL250 cells with no gross rearrangements as compared to the original BAC prep.

For recombinations (see below), electrocompetent cells were made from the EL250/BAC cells in the same manner except that before the 20 minute ice incubation step, the cultures were transferred immediately to 200 ml flasks and

placed in a 42°C shaker water bath for 15 minutes at 200 rpm to induce expression of recombination proteins.

BAC vectors were modified using homologous recombination in E. coli essentially as described (Mortlock et al. 2003) to contain a GFP-internal ribosome entry site (IRES) *lacZ*:Neo (β-geo) fusion cassette (pGIBG-FTet) inserted into the Bmp4 transcription unit. For simplicity, BAC RP23-145J23 was renamed 5' BAC and BAC RP23-26C16 was renamed 3' BAC. To generate the recombination cassette, 50-bp homology arms were designed to flank the start codon of *Bmp4* and additional sequence was added to the ends to allow for direct ligation into pGIBG-FTet as follows: for the 5' arm, 5' CTAGCTGCAGTGTTTATTTATTCTTTAACCTTCCACCCCAACCCCCTCCCCAG AGACACCTTAAT-3' (TOP), 5'-TAAGGTGTCTCTGGGGAGGGGGTTGGGGTGGAAGGTTAAAGAATAAATAAA CACTGCAG-3' (BOTTOM); for the 3' arm, 5'-CTAGTATGATTCCTGGTAACCGAATGCTGATGGTCGTTTTATTATGCCAAGTC CTCTCGAGC-3' (TOP), 5'-GGCCGCTCGAGAGGACTTGGCATAATAAAACGACCATCAGCATTCGGTTACC AGGAATCATA-3' (BOTTOM). Twenty micrograms of polyacrylamide gel electrophoresis (PAGE) purified oligonucleotides were annealed in 1X annealing buffer (0.1 M sodium chloride, 1 mM EDTA [pH 8.0], 10 mM Tris-HCL [pH 7.5]) to produce the double-stranded homology arms for direct ligation into pGIBG-FTet. The targeting cassette was isolated from pGIBG-FTet with a NotI, NheI double digest and gel purified prior to electroporation into recombination competent

bacterial cells containing the Bmp4 BACs. BAC recombination was performed as previously described (Lee et al. 2001). Recombinant clones were selected by tetracycline and chloramphenicol resistance and verified by restriction enzyme digestion with a rare-cutting enzyme followed by PFGE as described above. To remove the tetracycline resistance gene, verified clones were subjected to L-arabinose promoter-driven FLpe recombinase excision as previously described (Lee et al. 2001) (Mortlock et al. 2003). Finally, PFGE and fingerprint gel analysis was performed to verify the modified BACs (as described above). The following clones were selected for purification and pronuclear injection (see below): RP23-23C16-3.2 and RP23-145J23-3.1.

Bmp4 BAC Transgenic Mice

Purified BAC DNA constructs were used for pronuclear injections to generate founder mice and lines as previously described (Chandler et al. 2007). BAC DNA was harvested from 1 L of bacterial culture by alkaline lysis. First, a sample of glycerol-archived bacteria containing the modified BAC were streaked out on a LB plate containing chloramphenicol (CAM) and plates were incubated at 32°C overnight. A single isolated colony was used to inoculate a miniculture of 2 ml of LB containing CAM. Minicultures were incubated at 32°C overnight with agitation. Next, 2 ml of miniculture was used to inoculate 1 L of LB containing CAM in a 4 L flask. These large-scale cultures were incubated for approximately 20 hours at 32°C with agitation. The next day, each 1 L culture was centrifuged at 6,000 rpm in 2, 250 mL bottles (500 mL of culture per 250 mL bottle). To do

this, 250 mL of culture was added to the bottle and centrifuged. The supernatant was discarded leaving the pelleted cells in the bottle. The next 250 mL of culture was added to the same bottle with the pelleted cells and centrifugation was repeated in the same manner. Pelleted cells were incubated at -80°C for at least 30 minutes and then cells were resuspended in 50 mL of Solution I (50 mM D-(+)-Glucose, 25 mM Tris-HCL [pH 8.0], 10 mM EDTA, 50 µg/ml RNAse A) using a 10 mL pipet. Once cells appeared to be in solution, each bottle was agitated using a vortex. Then, 50 mL of fresh Solution II (0.2 M sodium hydroxide, 1% sodium dodecyl sulfate) was added to each 250 mL bottle and mixed gently for approximately 30 seconds by swirling and gently inverting bottles followed by incubation at room temperature for 5 minutes. Once cell lysis was complete, the solution was clear and not stringy. Next, 50 mL of cold Solution III (3 M potassium acetate [pH 5.5]) was added to each 250 mL bottle followed by gentle mixing by inversion and a 15 minute incubation on ice. Next, bottles were centrifuged at 10,000 rpm for 20 minutes at 4°C. The solid, white precipitate was discarded by filtering DNA supernatant through wet filter paper (Clontech, catalog #4062-1) into clean 250 mL bottles. An equal volume of molecular biology grade isopropanol at room temperature was added to supernatants followed by mixing by inversion and centrifugation at 10,000 rpm for 30 minutes at 4°C. DNA pellets were rinsed with 10 mL of 70% ethanol and centrifuged at 13,000 rpm for 10 minutes at 4°C. Ethanol was carefully poured off DNA pellets and DNA pellets were resuspended in 4 mL of TE (1M Tris-HCL [pH 7.5], 0.5 M EDTA [pH 8.0]. Finally, resuspended DNA pellets from each 250 mL bottle were

combined for a total final volume of 8 mL. BAC DNA was subsequently purified using a cesium chloride density centrifugation. First, 9.63 g of cesium chloride was added to a 15 mL conical tube followed by the 8 mL of DNA solution. Cesium chloride was dissolved into the DNA solution by gentle rocking at room temperature. Next, 0.8 mL of ethidium bromide solution (10 mg/mL) was added to the mixture. To remove solid precipitated material, tubes were centrifuged twice at 3,000 rpm for 5 minutes and DNA solution was transferred to a clean 15 mL tube in between spins. Finally, the DNA solution was added to Beckman OptiSeal 11.2 centrifuge tubes (Beckman, catalog #362181) using transfer pipets (Fisher, catalog #13-711-7M). Balanced tubes were centrifuged in a Beckman vTi65.1 vertical rotor at 65,000 rpm overnight at 16°C. The next day, supercoiled BAC DNA (lower band) was removed with a 21 guage needle in a volume of less than 1 mL. Ethidium bromide was removed from purified BAC DNA by at least six butanol extractions and BAC DNA was dialyzed against 3 L of microinjection buffer (10 mM Tris-HCl [pH 7.4], 0.15 mM EDTA [pH 8.0]) using 10,000-molecular-weight-cutoff Slide-A-Lyzer dialysis cassettes (Pierce, product #69570) followed by additional dialysis and concentration of BAC DNA with 30,000-molecular-weight-cutoff Centriprep centrifugal filter devices (Millipore, catalog #4306) to reduce the DNA solution final volume to less than 500 µL. Centriprep centrifugal devices were used according to the manufacturer's protocol (Millipore, catalog #4306) with one minor modification. To decrease the final volume of DNA solution, the last spins were performed at 5000 rpm versus 2800 rpm. BAC DNA samples were quantified as follows: digests with a rare-cutting restriction

enzyme (NruI) were analyzed by pulsed field gel electrophoresis in 1% agarose/ 0.5x Tris-Borate-EDTA buffer for 15 hours at 18°C (6 V/cm, 0.2-22s switch time) alongside known quantities of lambda DNA-HindIII digests as mass standards. To determine BAC DNA concentration, the gel was stained with ethidium bromide and Quantity One® Software was used to quantify BAC DNA bands by comparison to a standard curve of the lambda DNA-HindIII band intensities. The stock concentration of uncut BAC DNA was back calculated based on these estimates. Purified, circular BAC DNA was diluted to 1 ng/µL in microinjection buffer and used for pronuclear injections.

Bmp4 ^{lacZneo} Mice

Permission to use Bmp4 ^{lacZneo} mice (Lawson et al. 1999) was generously provided by Dr. Brigid Hogan. A mating pair was generously provided by Dr. Mark deCaestecker (Vanderbilt University) and Dr. David Frank (Vanderbilt University).

Genotyping

Bmp4 BAC transgenic mice were identified by a PCR-based genotyping strategy. Triplex PCR was performed on tail DNA samples using primers to detect *lacZ* in transgenic mice, the chloramphenicol resistance gene in transgenic mice, and *Gdf5* present in both transgenic and non-transgenic mice (control for PCR). PCR conditions were optimized by Laura Selenke, a former Research Technician in the lab. Primer sequences are as follows: for lacZ, 5'-

TTTCCATGTTGCCACTCGC -3' (forward), 5'- AACGGCTTGCCGTTCAGCA -3' (reverse); for chloramphenicol, 5'- GGAAATCGTCGTGGTATTCACTC-3' (forward), 5'- TCCCAATGGCATCGTAAAGAAC-3' (reverse); for *Gdf5*, 5'- TGGCACATCCAGAGACTAC -3' (forward), 5'- TGGAGAGAAATGAAGAGGC -3' (reverse). PCR conditions are as follows: 94°C for 5 min, 98°C for 5 sec, 94° C for 30 sec, 60°C for 1 min, 72°C for 40 sec (10 cycles); 94°C for 30 sec, 56°C for 1 min, 72°C for 40 sec (25 cycles); 72°C for 5 min. Copy number was estimated for founder mice and lines as described in Chapter IV. *Bmp4* BAC transgene integrity was analyzed by polymorphic marker analysis and copy number estimation as described in Chapter IV. *Bmp4* ^{lacZneo} mice were identified by visualizing *lacZ* expression in the hair follicles of tail snips after they were stained with Xgal (see below).

Transgene Expression Analysis

Bmp4 BAC transgene expression and *Bmp4* ^{lacZneo} expression was analyzed in embryos generated by test cross matings with transgenic males and wild-type Crl:CD1(ICR) females. Pregnant mice were sacrificed by CO_2 inhalation and their embryos were harvested for XGal staining to detect *lacZ* expression. Embryos were obtained at 9.5,12.5, and 15.5 days post coitus (dpc) for each line generated, allowing the expression of each line to be assayed throughout development. In brief, embryos were dissected into 1x phosphate buffered saline (PBS) on ice then fixed with 10% neutral buffered formalin at 4°C with agitation.

Embryos older than 14.5 dpc were bisected to allow for reagent penetration after fixation. Next, embryos were processed for XGal staining essentially as described (DiLeone et al. 1998) with two minor changes: 1) 0.6 mg/mL XGal was used and 2) embryos were stained overnight at room temperature with agitation.

Embryo Processing and Imaging

XGal stained embryos were staged into glycerol to promote clearing of the tissues. After XGal stained embryos were postfixed, they were put through a graded series of glycerol (EMD, catalog#356352) washes starting with 15% glycerol and proceeding through 30%, 50%, 70%, 90% and 100% glycerol. Glycerol solutions were made with 1X PBS. Each wash was performed at room temperature with agitation until embryos sank to the bottom of the vessel. The 100% glycerol washes were performed twice and embryos were stored in the final 100% glycerol wash.

Embryos were imaged with a digital camera on an Olympus SZX-ILLD2-100 stereomicroscope. Sections were imaged using a digital camera on an Olympus BX51 microscope.

Histology

To visualize *lacZ* expression on a cellular level, histological analysis was performed. XGal-stained, glycerol archived embryos were processed for paraffin sectioning. First, embryos in 100% glycerol were incubated in a 1:1 solution of glycerol:ethanol overnight with agitation. The following day, embryos were

washed in 30% glycerol/70% ethanol for one hour then placed in a graded series of ethanol solutions as follows: 70%, 80%, 90% ethanol for one hour each. This was followed by two washes in 100% ethanol for 30 minutes, then 30 minute incubations in Citrisolv (Fisherbrand) until embryos become very clear. Caution was used with incubation times in Citrisolv since overprocessing can result in brittle tissues. In addition, ethanol washes were not shortened due to the risk of not adequately removing water from tissues that can lead to paraffin infiltration problems. After the final Citrisolv incubation, embryos were incubated at 60°C for one hour in a 1:1 mixture of Paraplast® Plus Tissue Embedding Medium (paraffin):Citrisolv. Embryos were then incubated in 100% paraffin overnight at 60°C. The next day, embryos were incubated in a fresh change of 100% paraffin prior to embedding.

Embryos were sectioned at a thickness of 10 µm and mounted on Superfrost® Plus slides (Fisher). Slides were allowed to dry on a slide warmer overnight. Before slides were processed and stained, they were incubated at 60° C for 30 minutes to promote adherence of the tissues to the slides. Finally, sections were stained with either eosin or nuclear fast red (Vector Laboratories, catalog#H-3403) for approximately 5 minutes.

Results

Multiple lines were established for each GFP-IRES*lacZ*-BAC

Previous research has indicated that the proximal *Bmp4* promoter does not contain all the necessary elements to recapitulate endogenous expression in

mouse (Zhang et al. 2002) (Feng et al. 2002). Therefore, we employed a BAC-based strategy to test large segments of DNA containing *Bmp4* for regulatory activity. Each BAC was selected by using the UCSC Genome Browser to locate two separate BAC clones that both contained the *Bmp4* transcription unit and extended as far 5' or 3' to the gene as possible. Together, the 5' BAC (RP23-145J23) and 3' BAC (RP23-26C16) contain a 398 kb segment of mouse chromosome 14 including *Bmp4* (FIGURE 3.1). Each *Bmp4* BAC shares approximately 56 kb of common, overlapping sequence containing *Bmp4* (FIGURE 3.1). No other annotated genes are present in this BAC interval (FIGURE 3.1). However, a significant amount of cross-species conservation is present within the BAC interval (FIGURE 3.1), suggesting functional elements are present. In addition, the 5' BAC contains ECR 1 and 2 while the 3' BAC contains ECR 3 (FIGURE 3.1)(CHAPTER II).

Homologous recombination was used to insert a GFP-IRESβgeo cassette into the Bmp4 ATG start codon in each BAC. The predicted transcribed sequence for each BAC transgene includes *Bmp4* exon 1,2, and a small portion of exon 3 including the ATG, followed by GFP, IRES-βgeo, and the Sv40 polyadenylation signal (FIGURE 3.1). Therefore, GFP and lacZ (βgeo) are translated independently. This dual reporter cassette is functional as demonstrated by the presence of GFP fluorescence and *lacZ* staining in the same embryo (FIGURE 3.2).

Subsequent to pronuclear injection of each BAC transgene, multiple founder mice were identified with the 5' or 3' *Bmp4* GFP-*lacZ*-BAC transgene.

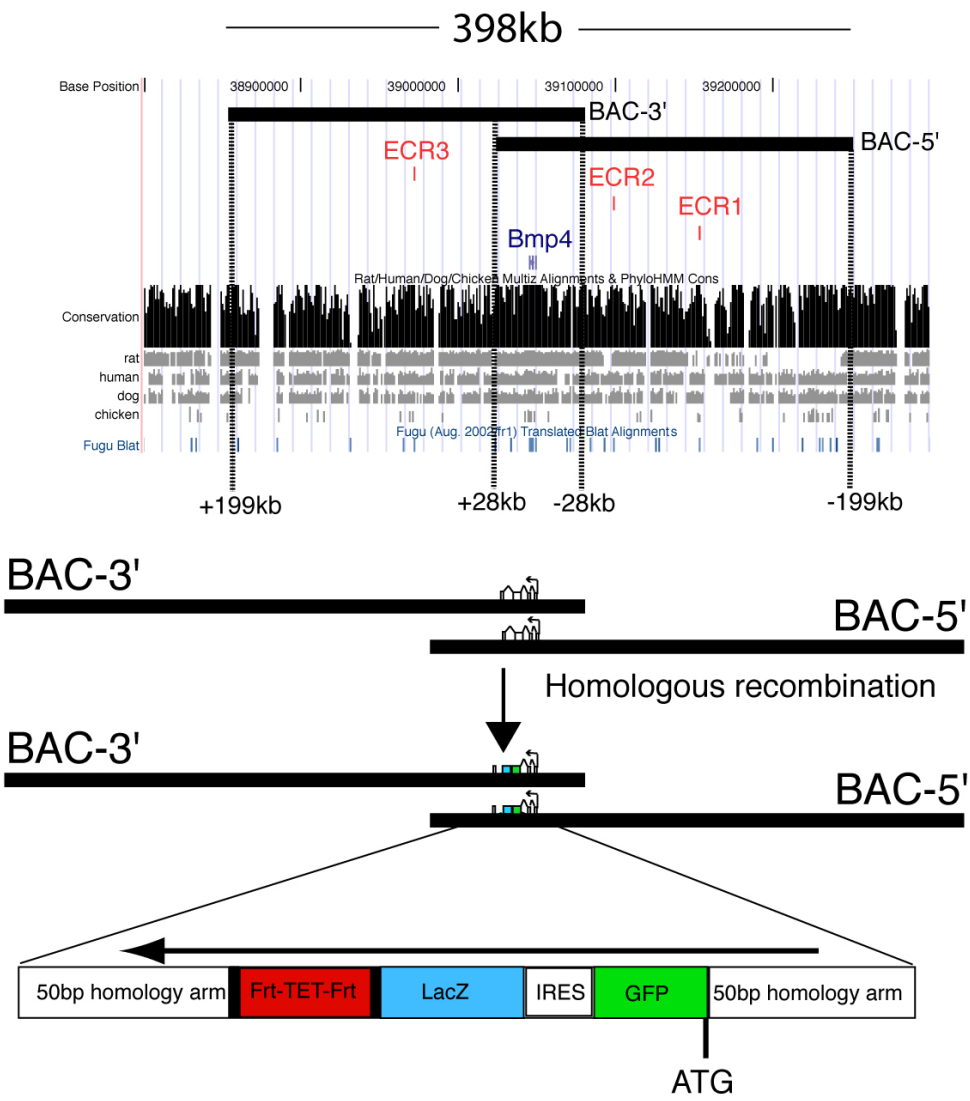

Figure 3.1. *Bmp4* BACs are modified into reporter transgenes. A 400 kb segment of mouse Chromosome 14 on the UCSC Genome Browser (Kent et al. 2002) depicts the location of each BAC used to generate reporter transgenes. Note, the 5' and 3' BAC each contain *Bmp4* in a 56 kb overlapping region. In addition, the 5' BAC contains ECR1 and 2, while the 3' BAC contains ECR3. Each BAC extends approximately 199 kb 5' or 3' to the *Bmp4* promoter. Homologous recombination techniques were employed to modify each BAC into dual GFP*lacZ* reporter transgenes.

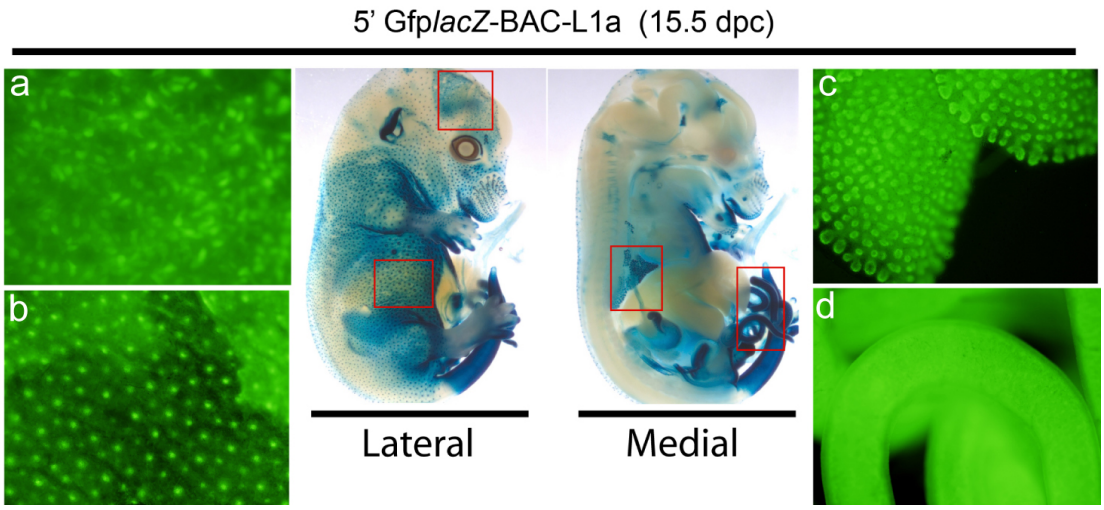

Figure 3.2. Dual GFPlacZ reporters function in *Bmp4* BACs. Shown here are images of GFP expression from a 15.5 dpc 5' *Bmp4* GFP*lacZ*-BAC embryo (L1a) prior to Xgal staining. Following GFP visualization, the same embryo was stained with Xgal to detect *lacZ* expression as depicted by the lateral and medial views of the bisected embryo. Both reporters exhibit robust expression as seen in multiple structures such as the (a) calvaria, (b) pelage hair follicle placodes, (c) lung epithelium, and (d) gut.

Embryos were generated at three different developmental stages (9.5, 12.5, 15.5 dpc) for all fertile lines and assayed for *lacZ* activity. This allowed us to assess *lacZ* reporter expression in a select number of embryos for all lines before looking at expression in more detail throughout development. Nine founders were identified for the 5' *Bmp4* GFP-*lacZ*-BAC. Two lines (L1a, L83) had robust transgene expression as demonstrated by XGal staining and were identical in their expression patterns (FIGURE 3.3). In addition, polymorphic marker analysis (L1a, L83), Southern blot analysis (L1a), and copy number estimation (L1a, L83) suggests both lines most likely contain at least one copy of an intact BAC transgene (CHAPTER IV). In addition to L1a and L83, L69 demonstrated strong *lacZ* expression in patterns that were reproduced by the previous two lines at 12.5 dpc. Unfortunately, L69 failed to generate fertile transgenic progeny after a year of breeding and screening efforts. In addition, this line appeared to be mosaic in the germline since the rate of transgenesis was well below expected Mendelian ratios. Because the expression levels in L69 were robust and it was a high copy line (CHAPTER IV), emphasis was placed on obtaining as much data as possible from this transgenic male. Therefore, data from this line was generated with one 12.5 dpc transgenic embryo (see FIGURE 3.11). Notice, the expression patterns seen in L69 closely recapitulated a subset of endogenous Bmp4 expression patterns seen in the *Bmp4* knock-in line (see FIGURE 3.11).

Four of nine lines exhibited low (L12) or undetectable (L52, L73, L90) expression levels. One founder contained two independent transgene insertions as demonstrated by: 1) a higher than expected rate of transmission to progeny

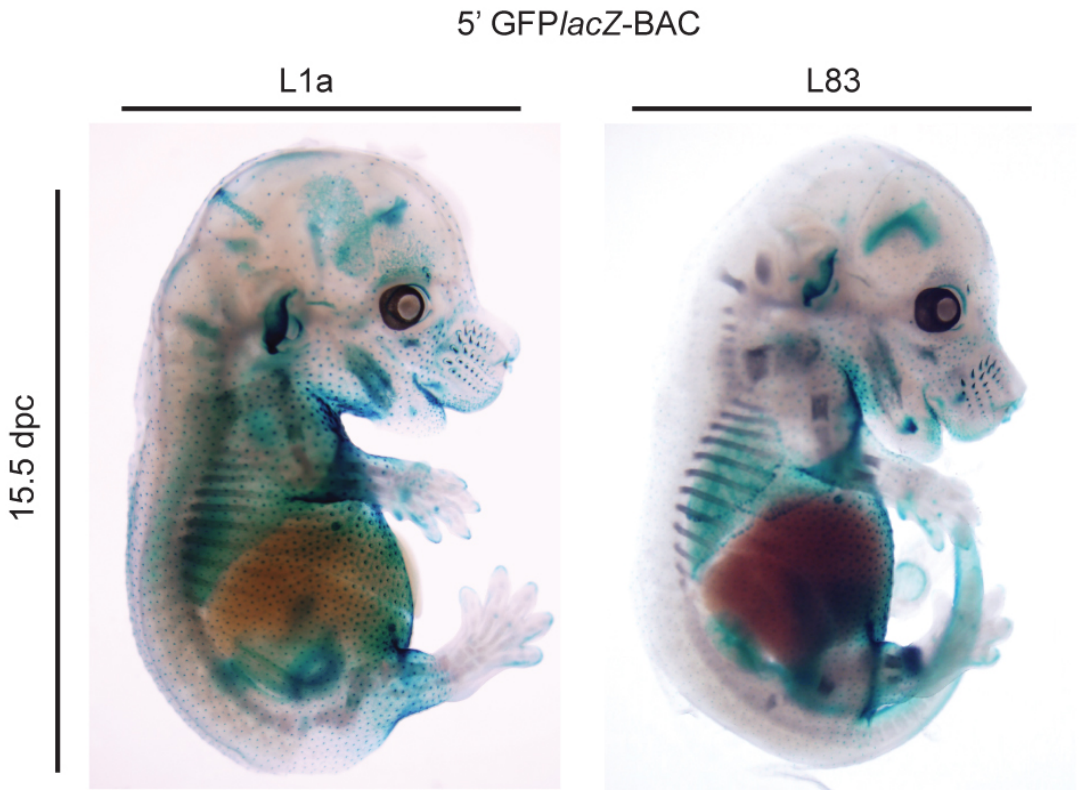

Figure 3.3. Expression patterns are reproducible in two independent 5' *Bmp4* GFP*lacZ*- BAC lines. Xgal-stained embryos at 15.5 dpc from L1a (left) and L83 (right) exhibit *lacZ* expression in similar tissues.

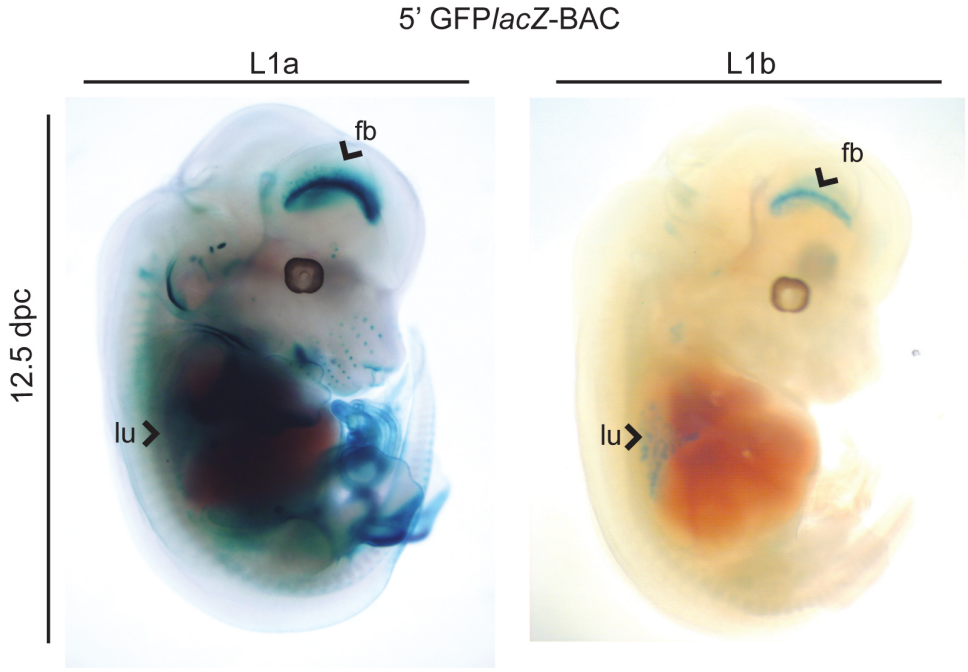

Figure 3.4. Expression patterns from the two independent transgene insertions derived from the 5' *Bmp4* GFP*lacZ*- BAC founder L1. Xgal-stained embryos at 12.5 dpc from L1a (left) and L1b (right) show dramatic differences in expression. L1a exhibits expression in numerous tissues including the lung and forebrain, whereas L1b has expression in the lung and forebrain alone (arrows). In addition, expression appears to be much weaker in lung/forebrain of L1b and much stronger in lung/forebrain of L1a. Subsequent polymorphic marker and Southern blot analysis (see Chapter IV) suggest L1a contains at least one intact transgene.

(75%) and 2) independent segregation of each insertion. A line generated by this founder (L1b) carries only one of the insertions and appears to contain a fragmented transgene since expression is only seen in a limited number of sites (eg. forebrain, lung) (FIGURE 3.4, arrowheads), while the other insertion directs multiple sites of expression in L1a. Southern blot and polymorphic marker analysis (see CHAPTER IV) indicate L1a contains at least one copy of an intact transgene (see FIGURE 4.9). However, further analysis such as polymorphic marker genotyping or Southern blots were not performed on L1b to further support this hypothesis. In total, three lines (L1a, L69, L83) were used to perform data analysis presented in Figure 3.8a. As stated previously, one embryo was generated at 12.5 dpc from L69 and no embryos were generated at 9.5 or 15.5 dpc. Therefore, this line contributed data from 12.5 dpc in Figure 3.8a. Overall, each each expression pattern was replicated in at least two independent 5' BAC lines (see FIGURE 3.8a).

In addition, eleven founder mice were identified with the 3' GFP-*lacZ*-BAC. Five of eleven 3' BAC lines (L19, L37, L45a, L46, L57) exhibited moderate-to-robust reporter expression, high copy number (CHAPTER IV), and reproducible expression patterns, suggesting these lines contained intact transgenes (FIGURE 3.6). Seven of the eleven lines were analyzed for integrity by polymorphic marker analysis and copy number estimation (L37, L44, L45a, L46, L50, L57, L65) (CHAPTER IV). Of these lines, five had intact transgenes (L37, L45a, L46, L50, L57) and two were fragmented (L44,L65) according to polymorphic marker analysis (CHAPTER IV). Although marker analysis

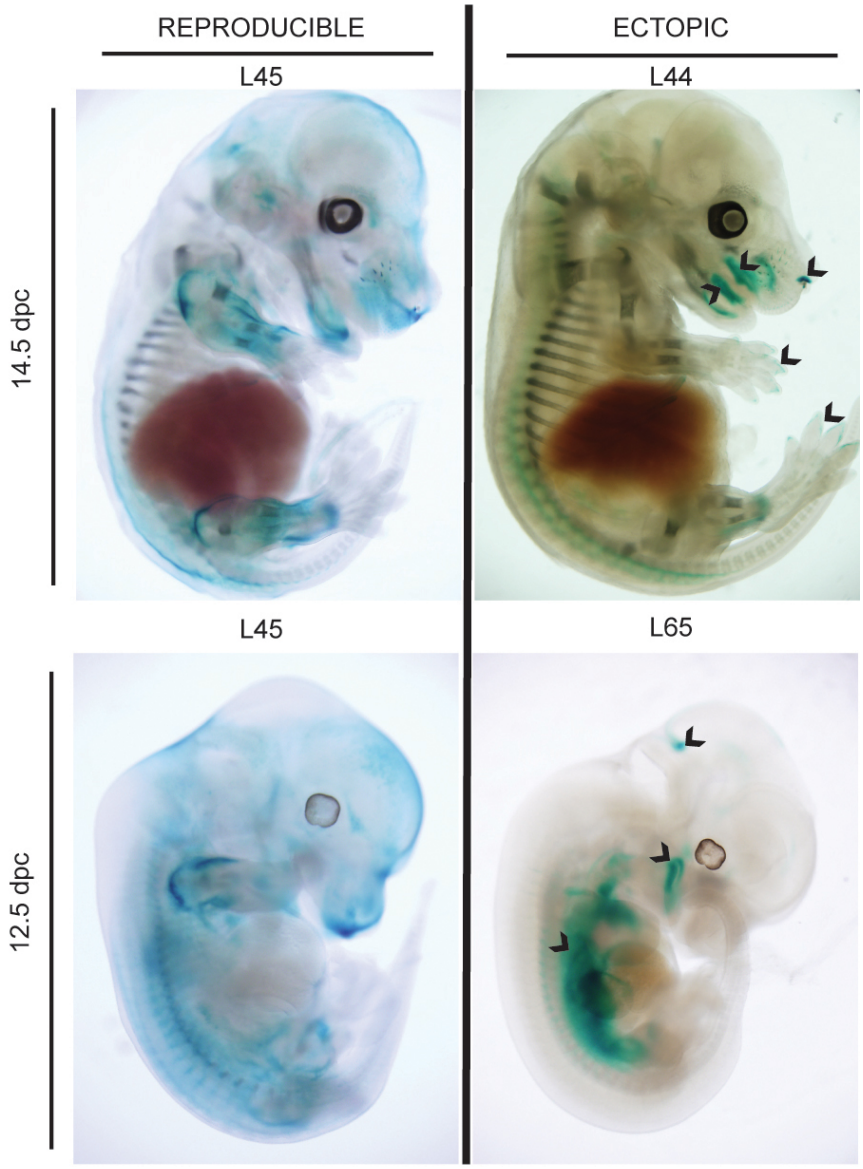

Figure 3.5. Two of nine 3' *Bmp4* GFP/*lacZ*-BAC lines exhibit ectopic reporter expression. At 14.5 dpc, Xgal-staining reveals L44 (right) has ectopic expression in the mouth, nose and sides of the digits (arrowheads). These expression patterns are not seen in the representative 3' BAC L45a (left) or in age-matched *Bmp4* knock-in embryos (see FIGURE 3.8). Likewise, L65 shows ectopic expression in the brain, corners of the mouth, and the internal midline unlike expression seen in the representative 3' BAC L45a.

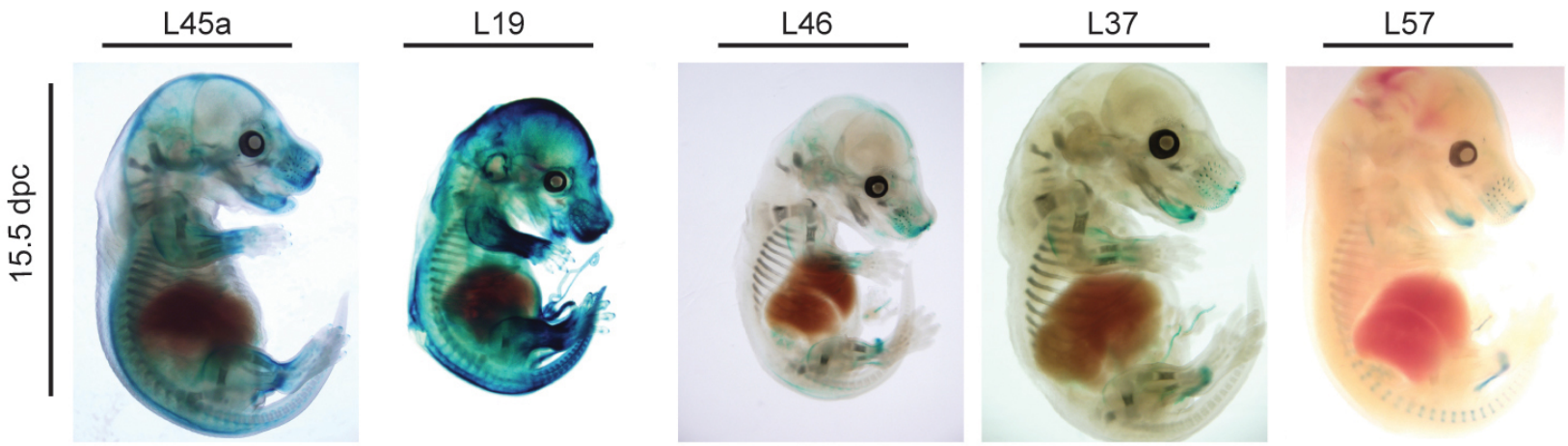

Figure 3.6. Expression patterns are reproducible in five independent 3' *Bmp4* GFP*lacZ*-BAC lines. Xgal-stained embryos at 15.5 dpc exhibit varying levels of *lacZ* expression in the same tissues. For example, each line displays expression in the ventral ribs, dura mater, proximal limb mesenchyme, craniofacial mesenchyme, and whisker hair shaft. Although each line also has *lacZ* expression in the umbilical artery, kidney and genital tubercle, these structures are either not visible or obstructed by the hindlimb in some lines. In addition, L37 has very faint expression in the vertebral column and dura mater that is not obvious in the photograph, but is obvious in the remaining lines. Also, all lines exhibit *lacZ* expression in the digit tips, although it is difficult to see in some of these images. Reporter expression appears to be associated with transgene copy number (see Chapter IV).

suggested L50 contained at least one full-length transgene, reporter expression was not detectable. Both lines with fragmented transgenes (L44, L65) displayed ectopic expression patterns (FIGURE 3.5). For example, L44 had expression in the mouth, nose, and sides of the digits (FIGURE 3.5, arrowheads) that were not present in any of the other 3' GFP-*lacZ*-BAC lines or the *Bmp4* knock in line. L65 exhibited *lacZ* expression in the brain, mouse, and internal midline that was not present in any of the other 3' GFP-*lacZ*-BAC lines or the *Bmp4* knock in line (FIGURE 3.5, arrowheads). Note, the 5' and 3' GFP-*lacZ*-BAC lines that were used in our analysis and compilation of expression patterns (FIGURE 3.8A) did not exhibit ectopic expression patterns. Interestingly, one line (L19) exhibited a defect in the frontal bones of the skull (FIGURE 3.7) that segregated with the transgene. No other lines displayed any noticeable physical abnormalities. In addition, one founder maintained two independent insertion sites for the 3' GFP-l*acZ*-BAC transgene that segregated independently in subsequent breedings (L45a, L45b) (CHAPTER IV). In sum, five 3' BAC lines (L19, L37, L45a, L46, L57) were used for data analysis.

Bmp4 lacZ-BAC Transgenes Direct Multiple Unique Sites of Expression Suggesting Multiple Long-Range Enhancers are Present within the BAC Interval

Previous studies in mouse and fish suggest many *Bmp4 cis*-regulatory elements act at a great distance from the promoter (Feng et al. 2002) (Zhang et al. 2002) (Shentu et al. 2003). Likewise, other Bmp family members, including *Bmp4's* close homolog, *Bmp2*, have been shown to utilize long-range regulatory elements to direct their complex developmentally-regulated expression patterns

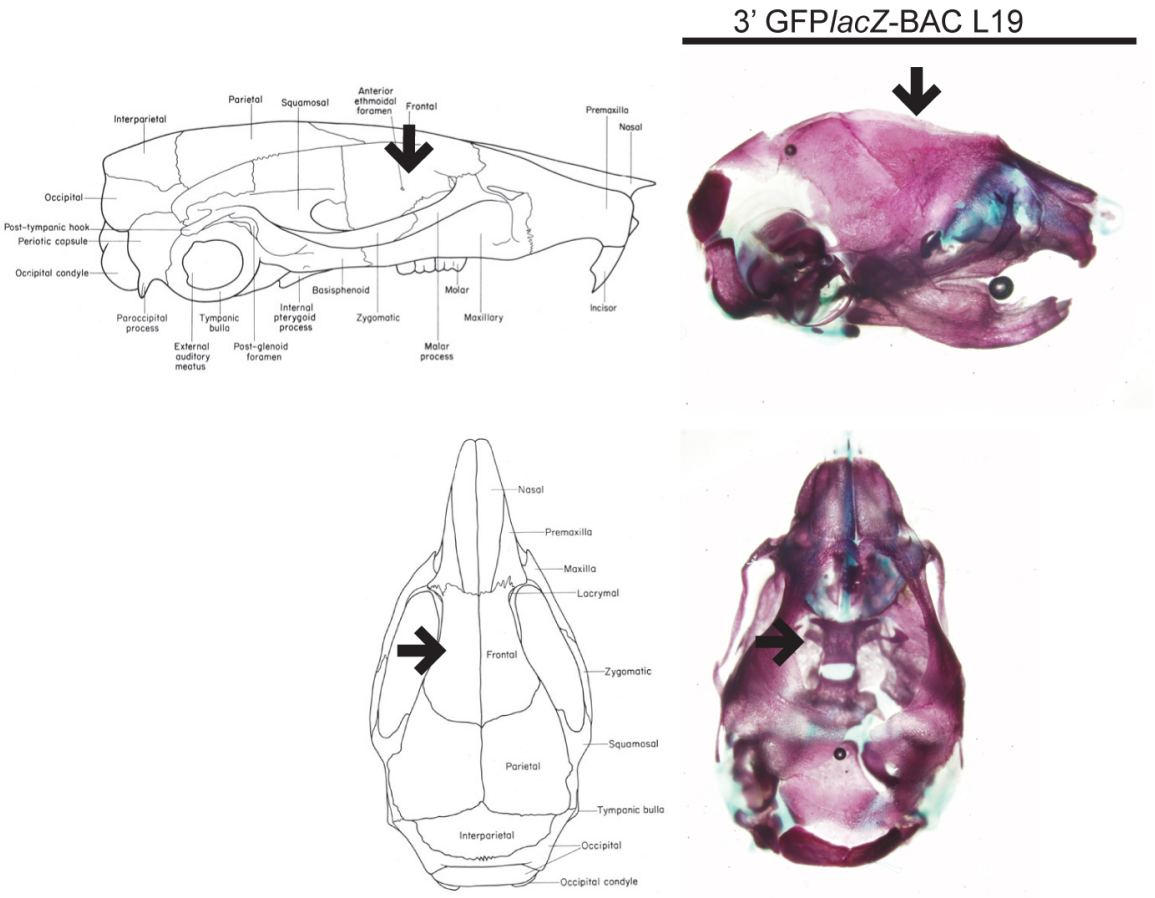

Figure 3.7. A 3' GFP*lacZ*-BAC line (L19) is missing the frontal skull bone. Alizarin red (bone) and alcian blue (cartilage)-stained skull of a 3' GFP*lacZ*-BAC L19 newborn pup reveals the frontal skull bone (arrows) failed to develop. Adapted from (Cook 1965).

(Mortlock et al. 2003) (Chandler et al. 2007) (DiLeone et al. 2000). Studies have shown modified bacterial artificial chromosome (BAC) reporter transgenes can successfully parse gene deserts for *cis*-regulatory activity (Mortlock et al. 2003) (Chandler et al. 2007). Therefore, two overlapping *Bmp4* BACs were modified into reporter transgenes. To assess the *cis*-regulatory activity of each *Bmp4* GFP-*lacZ*-BAC transgene, embryos from multiple lines for each BAC were stained with XGal to detect *lacZ* activity. Multiple lines were generated to ensure independent replication of each data point. Prior to data analysis, each line was assessed for *lacZ* activity at three different developmental stages (9.5, 12.5, 15.5 dpc). These stages were chosen because they capture both the onset and the completion of organogenesis. Since we hypothesize *Bmp4* maintains multiple long-range *cis*-regulatory elements to impart developmentally regulated gene expression, analysis of these stages should account for many *Bmp4* expression patterns that occur in development. After embryos were collected from each line at three different developmental stages (9.5, 12.5, 15.5 dpc) and assayed for *lacZ* activity, a representative line was chosen for each BAC. Strength of *lacZ* expression varied amongst lines. Further analysis suggested that increased copy number correlated with increased *lacZ* expression (see CHAPTER IV). Representative lines (5' GFP-IRES*lacZ* BAC L1a, 3' GFP-IRES*lacZ* BAC L45a) were selected for their robust *lacZ* expression as well as reproducibility of expression patterns in other lines bearing the same transgene. Next, embryos were generated from each representative line for each developmental day starting at 6.5 dpc and ending with 15.5 dpc to obtain a more detailed view of

Bmp4 BAC transgene expression throughout the majority of mouse development. In addition, age-matched embryos were generated from the *Bmp4* knock in line and *lacZ* expression was compared to BAC transgene expression. Expression data was gathered from embryos at 9.5, 12.5, and 15.5 dpc from each 5' and 3' *Bmp4* BAC line and compiled. Lines were examined for *lacZ* expression in embryos generated at each developmental stage and compared to age-matched embryos generated by the *Bmp4* knock in line. Expression patterns that reflected endogenous *Bmp4* expression as determined by the *Bmp4* knock in line were scored as present or absent in each line and this is summarized in Figure 3.8. Lines with undetectable *lacZ* activity and/or a fragmented transgene as identified by polymorphic marker analysis (see CHAPTER IV) were not included in the compiled data (FIGURE 3.8). Since each BAC shared a common overlapping region of approximately 56 kb (FIGURE 3.8), we expected to see some patterns of expression that were common to both BAC transgenes. This would indicate that proximal enhancers reside in the common, overlapping domain. In addition, each BAC contained approximately 171 kb of unique genomic sequence (FIGURE 3.8). Therefore, we expected each BAC transgene would also direct its own unique set of expression patterns, indicating long-range enhancers reside 5' or 3' to *Bmp4*.

Each *Bmp4* BAC directed a common set of expression patterns. For example, both the 5' and 3' BAC drove *lacZ* expression in the whisker hair shaft at 15.5 dpc (see FIGURE 3.12c, i). Likewise, each BAC directed *lacZ* expression in the genital tubercle (FIGURE 3.9), digit tips (FIGURE 3.9) and dorsal root

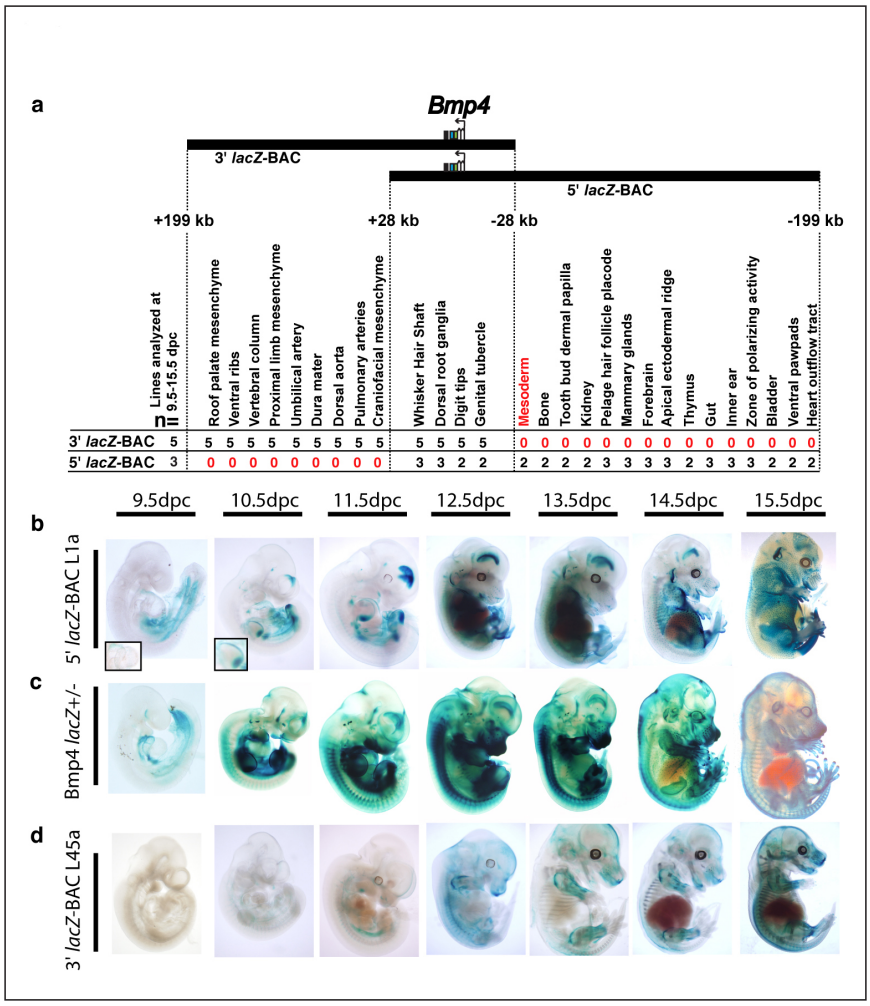

Figure 3.8. *Bmp4* BAC transgenes direct some common sites of expression and multiple unique sites of expression during embryonic development. (a) Two BAC clones are modified to contain GFP*lacZ* reporters in the ATG of *Bmp4* exon 3. *Bmp4* is on the minus strand here, as indicated by an arrow pointing to the left. GFP (green box) and *lacZ* (blue box) are inserted in exon 3. Together, the 5' and 3' BACs cover nearly 400 kb (+199 kb, -199 kb) encompassing mouse *Bmp4* (not drawn to scale). Below each BAC transgene are the anatomical sites where *lacZ* was expressed throughout development (9.5,12.5, and 15.5 dpc). A total of five lines were examined for the 3' BAC and three lines for the 5' BAC. Listed below each anatomical site is the number of lines that exhibited *lacZ* expression in that site. (b-d) Embryos generated from the representative (b) 5' BAC L1a and (d) 3' BAC L45a, as well as age-matched embryos from the (c) *Bmp4* knock-in line are stained with Xgal to detect *lacZ* expression throughout embryonic development. Note, the 5' BAC 9.5 dpc embryo has an inset image of the outflow tract in the heart and the 10.5 dpc embryos has an inset image of the forelimb to better visualize the expression.

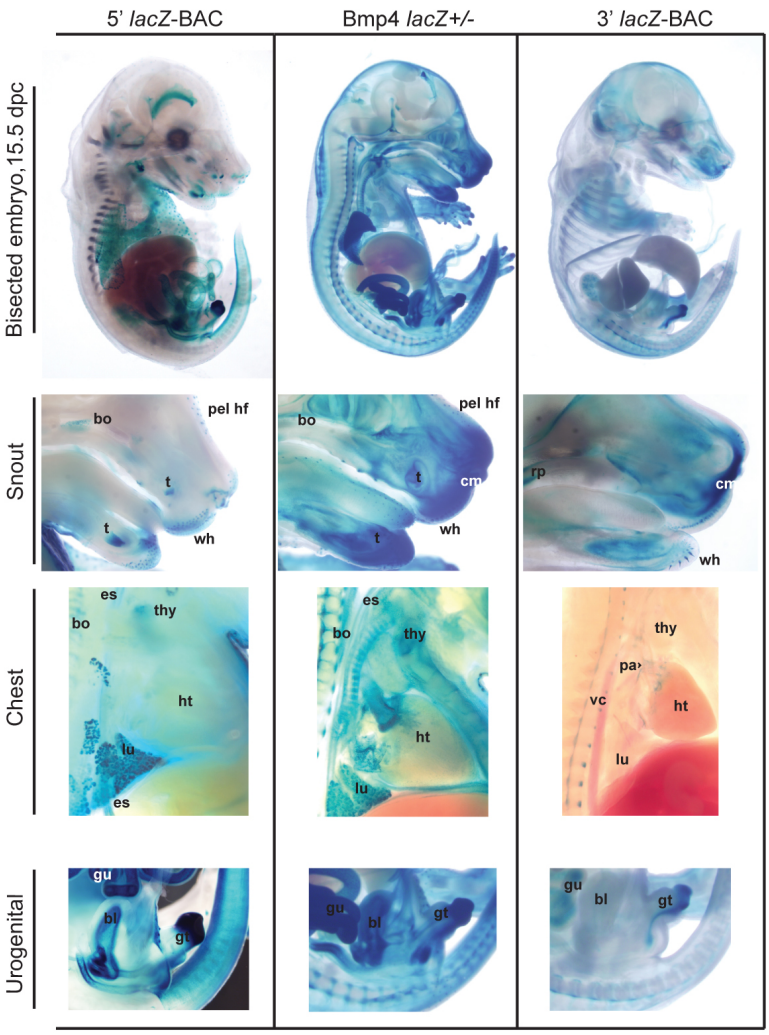

Figure 3.9. Expression patterns in *Bmp4* BAC embryos reflect endogenous *Bmp4* expression. Xgal-stained, bisected 15.5 dpc embryos (medial view) from the 5' BAC L1a (left), *Bmp4* knock-in (middle) and 3' BAC L45a (right) allow comparison of transgene-directed and *Bmp4*-directed *lacZ* expression patterns. Enlarged images of the bisected snout region (second row) indicate the 5' BAC directs expression in the tooth (t), whisker (wh), bone (bo), pelage hair follicle placodes (pel hf), similar to the *Bmp4* knock-in embryo. The 3' BAC also directs expression in the whisker (wh). In addition, the 3' BAC directs expression in the roof palate (rp) and craniofacial mesenchyme (cm). Enlarged images of the bisected thoracic cavity (third row) show the 5' BAC directs expression in the lung (lu), thymus (thy) (located behind tissue), esophagus (es) and bone (bo), while the 3' BAC directs expression along the vertebral column (vc) and pulmonary artery (pa). Enlarged images of the bisected posterior of each embryo indicates the 5' BAC directs expression in the bladder (bl), gut (gu) and genital tubercle (gt), while the 3' BAC directs expression in the genital tubercle (gt).

ganglia (FIGURE 3.8, 12.5 dpc). In contrast, sometimes each BAC directed expression in the same tissue, but in different patterns. For example, each BAC transgene directed expression in the kidney at 15.5 dpc. However, the 5' BAC directed expression in both mesenchyme and epithelial cells in the kidney (see FIGURE 3.12b), while the 3' BAC directed expression solely in epithelial cells (see FIGURE 3.12h). Therefore, epithelial cell expression in the kidney is most likely controlled by an element common to both BACs, while mesenchymal cell expression in the kidney is unique to the 5' BAC only. Overall, these sites of expression reflect endogenous *Bmp4*, as demonstrated by the knock-in mouse (FIGURES 3.8 and 3.9).

The 5' *Bmp4* BAC directed numerous sites of expression that were never seen in any of the 3' *Bmp4* BAC lines, but were present in the *Bmp4* knock-in line. After gastrulation commenced (7.5 dpc), the 5' GFP*lacZ*-BAC drove expression in the extraembryonic mesoderm (FIGURE 3.10). To confirm the expression pattern matched that of endogenous *Bmp4*, *Bmp4* knock in embryos were generated at 7.5 dpc and stained with XGal (data not shown). The extraembryonic mesoderm was devoid of *lacZ* expression in the representative 3' BAC line (L45a) (data not shown), indicating regulatory element(s) located between 28-199 kb 5' to *Bmp4* directed *Bmp4* expression in the extraembryonic mesoderm. Likewise, the 5' GFP*lacZ*-BAC drove expression in the posterior lateral plate mesoderm as well as the foregut (FIGURE 3.10), and outflow tract of the developing heart at 9.5 dpc (FIGURE 3.8b, inset). However, the 3' BAC failed to direct these patterns of expression at 9.5 dpc (FIGURE 3.8d). By 10.5

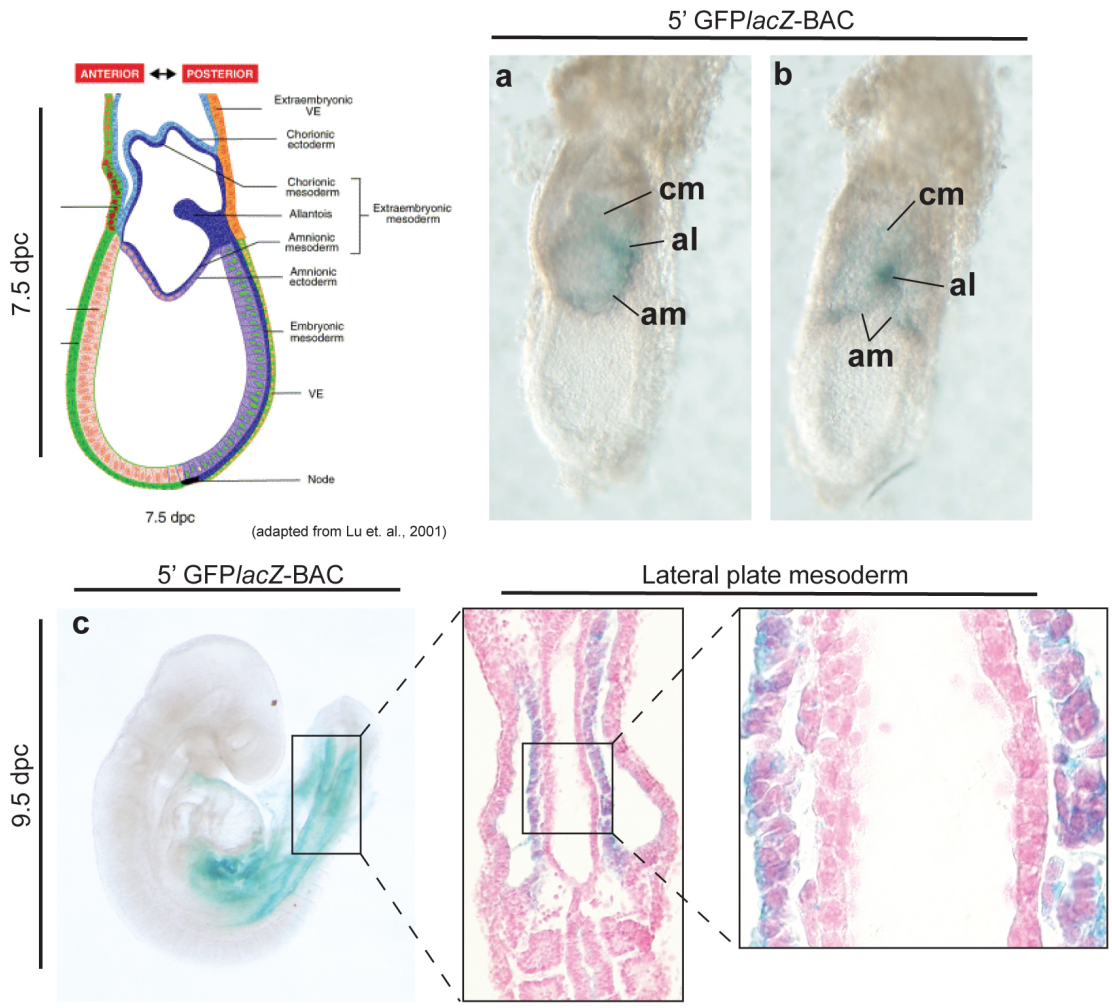

Figure 3.10. 5' BAC directs expression in extraembryonic and lateral plate mesoderm. Top left, cartoon depicting the anatomy of a 7.5 dpc mouse embryo. (a and b) At 7.5 dpc, the 5' BAC directs expression only in the extraembryonic mesoderm. (c) By 9.5 dpc, the 5' BAC directs expression in the lateral plate mesoderm as seen in histological sections (inset). cm, chorionic mesoderm. al, allantois. am, amniotic mesoderm. Top left cartoon adapted from (Lu et al. 2001).

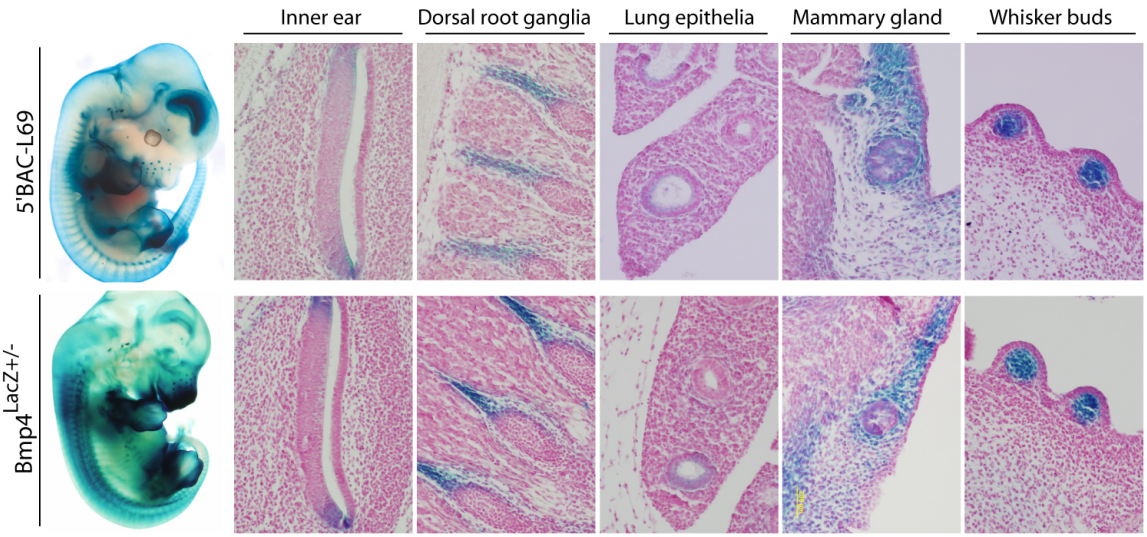

Figure 3.11. Expression directed by 5' BAC transgene reflect endogenous *Bmp4* expression patterns. Xgal stained embryos from the 5' BAC L69 and *Bmp4* knock-in line at 12.5 dpc are serially sectioned and counterstained to visualize cellular localization of *lacZ* staining. Shown here are a sample of expression patterns including the inner ear, dorsal root ganglia, lung epithelia, mammary gland and whisker buds.

dpc, *lacZ* expression was detected in the forebrain, apical ectodermal ridge (AER), and a posterior zone in the limb bud in 5' BAC embryos (FIGURE 3.8b, inset). However, 3' BAC age-matched embryos were devoid of these expression patterns (FIGURE 3.8d).

Later in development, the 5' *Bmp4* BAC drove expression of GFP/*lacZ* in the distal epithelium of the branching lung at 12.5 (FIGURE 3.11) and 15.5 dpc (see FIGURES 3.2, 3.9, 3.12a), as well as in the pelage hair follicles in a dramatic spotted pattern (see FIGURE 3.2). In addition, the 5' BAC alone directed expression in tooth (FIGURE 3.12e), bladder, ventral pawpads, forebrain, bone (FIGURE 3.12f), kidney mesenchyme (FIGURE 3.12b), thymus, stomach and gut (FIGURE 3.2, 3.12d, and see FIGURE 3.9) at 15.5 dpc. The transgene driven expression was compared to the *Bmp4* knock-in mouse to verify expression patterns were not ectopic (see FIGURES 3.8 and 3.9). Histological sections through 5' BAC and *Bmp4* knock in embryos further demonstrate the replication of endogenous expression by the transgene driven expression patterns (FIGURES 3.10 and 3.11).

Multiple *lacZ* expression patterns were also found only in the 3' GFP*lacZ*-BAC embryos, but never in the 5' GFP*lacZ*-BAC embryos. Reporter expression is first noted at 10.5 dpc as thin stripes in a segmental pattern along the dorsal region of the 3' BAC embryo (see FIGURE 3.8). By 12.5 dpc, *lacZ* expression is detected in the craniofacial and proximal limb mesenchyme (3.12k, see FIGURE 3.8). After the bulk of organogenesis is completed (15.5 dpc), *lacZ* expression was seen in the vertebral column (FIGURE 3.12m), dura mater (FIGURE 3.12j),

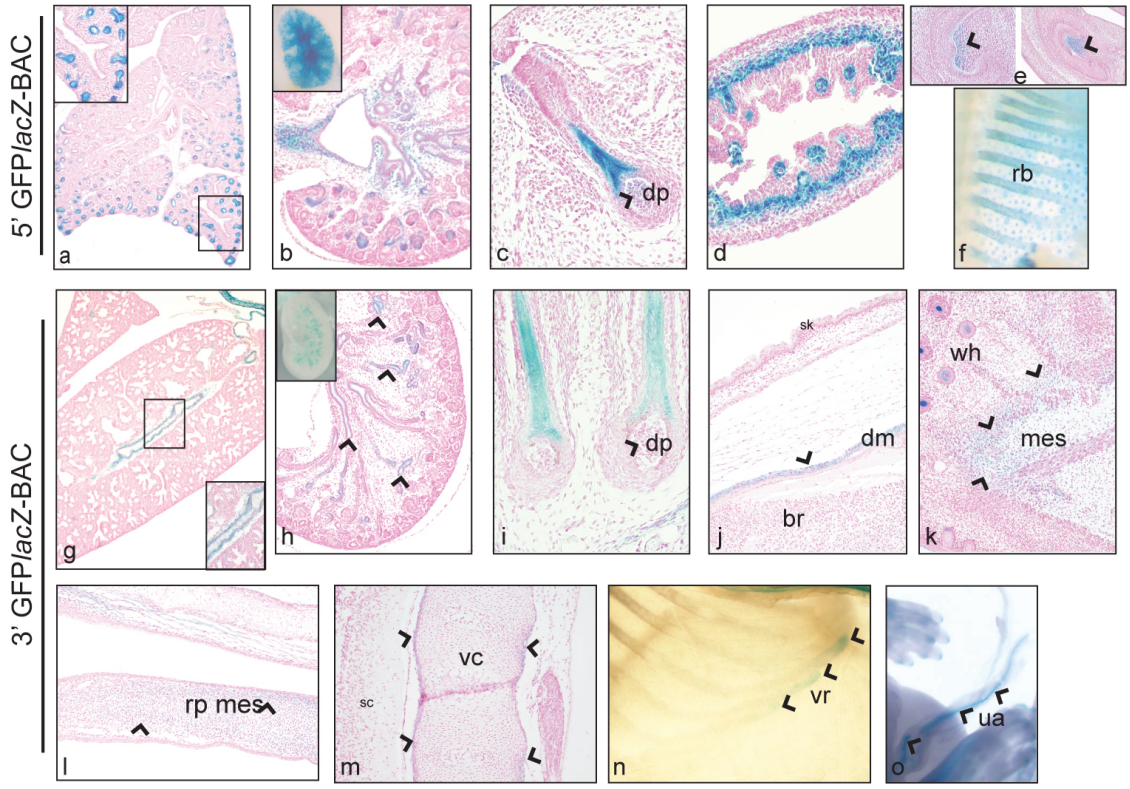

Figure 3.12. Cellular localization of *lacZ* expression in 5' and 3' BAC lines. (a-e) Xgal-stained embryos at 15.5 dpc from the 5' BAC L1a reveal 5' BAC-directed expression in the following structures: (a) lung epithelia (b) kidney epithelia and mesenchyme, (c) whisker hair shaft and dermal papilla (dp), (d) gut mesenchyme, (e) upper tooth dermal papilla (left), lower tooth dermal papilla (right). (f) Whole-mount image of 15.5 dpc 5' BAC embryo showing *lacZ* expression in the pelage hair follicle placodes as well as the rib bones (rb). (g-m) Xgal-stained embryos at 15.5 dpc from the 3' BAC L45a reveal 3' BAC-directed expression in the following structures: (g) pulmonary artery in lung, (h) kidney epithelium, (i) whisker hair shaft, (j) dura mater (dm), (k) craniofacial mesenchyme (mes) and whisker hair shaft (wh), (l) roof palate mesenchyme (rp mes), and (m) vertebral column (vc). (n-o) Whole-mount images of 15.5 dpc 3' BAC embryo showing *lacZ* expression in the (n) ventral ribs (vr) and the (o) umbilical artery (ua). (sc=spinal cord)

ventral ribs (FIGURE 3.12n), roof plate mesenchyme (FIGURE 3.12l), umbilical artery (FIGURE 3.12o), dorsal aorta and pulmonary arteries (see FIGURE 3.9). These expression patterns were also present in the *Bmp4* knock in embryos (see FIGURE 3.9). Therefore, we hypothesize regulatory elements for expression patterns seen in 3', but not 5', BAC embryos are located in the approximately 171 kb interval 3' to *Bmp4* (see FIGURE 3.8).

Interestingly, several sites where *Bmp4* is expressed did not stain with Xgal in lines generated from either BAC transgene. For example, neither BAC transgene drove expression in extraembryonic ectoderm (Lawson et al. 1999), eye (see FIGURE 3.11), trachea (see FIGURE 3.9), or anterior limb bud (see FIGURE 3.11). Therefore, together, both BAC transgenes failed to direct expression in all known sites of *Bmp4* expression, suggesting regulatory elements exist beyond the confines of the BAC intervals tested.

Discussion

Developmentally regulated genes, like *Bmp4*, have been shown to maintain a repertoire of *cis*-regulatory elements dispersed throughout a vast genomic region (Sandelin et al. 2004) (Plessy et al. 2005) (Gomez-Skarmeta et al. 2006) (Chandler et al. 2007) (DiLeone et al. 1998) (Wunderle et al. 1998) (Kimura-Yoshida et al. 2004) (Lettice et al. 2003) (Lettice et al. 2002) (Mortlock et al. 2003) (Nobrega et al. 2003). The widespread nature of a gene can present a challenge for identification of *cis*-regulatory elements using conventional plasmid-sized constructs. Therefore, we modified BAC vectors into reporter trangenes to

test large segments of genomic DNA for transcriptional activity. Using this approach, our data suggests a distant 171 kb 3' interval and a distant 171 kb 5' interval flanking *Bmp4* harbor unique regulatory functions. Furthermore, our data strongly suggests multiple, long-range regulatory elements direct *Bmp4* expression.

Using these methods, we were able to establish multiple lines for both the 5' and 3' BAC transgene constructs. We found the presence or strength of expression varied between lines established from the same construct. This is examined and discussed in great detail in Chapter IV. Briefly, lines with high copy number tended to have more robust reporter expression. High copy number may be necessary to overcome position effects such as silencing when the transgene randomly integrates into the genome. Regardless, all but two lines with visible reporter expression had expression patterns that were reproducible. These exceptions (L44, L65) were also shown to contain only a portion of the original BAC transgene (see FIGURE 3.5). Therefore, the ectopic expression patterns that were seen could be due to a lack of repressor elements present in the missing portion of the BAC transgene. Alternatively, the transgene may have integrated into a genomic locus with other regulatory elements that are able to initiate a transcriptional response from the *Bmp4* promoter and direct expression in the ectopic sites. Overall, lack of expression or ectopic expression was overcome by generating numerous lines for each BAC to ensure reproducibility of expression. A slight difference in *lacZ* expression between the BAC transgenic embryos versus the *Bmp4* knock in embryos is due to the nuclear localization

signal present in the knock in allele. Therefore, *Bmp4* knock in embryos will exhibit *lacZ* expression in the nucleus of *Bmp4* expressing cells rather than the cytoplasm leading to a more restricted expression pattern in the cells of interest. For example, histological sections generated from the BAC transgenic embryo exhibit *lacZ* expression in more diffuse patterns, whereas histological sections obtained from the *Bmp4* knock-in embryos exhibit nuclear-restricted expression patterns (see FIGURE 3.11).

Interestingly, a 3' BAC line exhibited a skull defect that segregated with the transgene. This line (L45a) had the highest copy number of all lines generated (~100 copies) (see CHAPTER IV). Although we did not perform the necessary experiments to flush out the cause of this defect, we can propose two hypotheses that might explain the absence of frontal bones in the transgenic mice. One hypothesis is that the transgene array inserted into the locus of another gene, thereby disrupting its function. If there were actually 100 physical copies of the BAC transgene, an array spanning 20 megabases (Mb) would be present in the genome. Not only could an array of this size lead to the physical disruption of a gene involved in frontal bone morphogenesis, but it may also sequester transcription factors necessary for frontal bone development thereby depleting the cell of that transcription factor. For example, the *Msx1/2* double knockout mouse completely lacks the frontal bone (Han et al. 2007). The high copy 3' BAC L45a phenocopies the *Msx1/2* knockout. Cells that are normally fated to become frontal bone cells may be deficient in *Msx1/2* because the 3' BAC is present in so many copies that it prevents correct regulation of another

gene or gene(s) that must bind Msx1/2 to elicit transcription in those cells, allowing their fate to be determined. Therefore, the gene(s) that needs to be transcribed to allow frontal bone development are not transcribed due to a lack of necessary transcription factors that are “soaked up” by the transgene array. While we did not examine if 3’ BAC expression overlaps *Msx1/2* expression in the skull, this may be an area of future investigation.

We show evidence that numerous *cis*-regulatory elements are located at least 28 kb 5’ or 3’ to *Bmp4*. Therefore, *Bmp4* maintains a widespread and complex *cis*-regulatory landscape similar to its ancestral gene, *dpp*. *Bmp4* utilizes multiple long-range *cis*-regulatory elements located both 5’ and 3’ to the gene to impart spatiotemporal expression throughout development. Some sites where *Bmp4* is expressed are unstained by Xgal in either the 5’ or 3’ BAC transgenic lines. This may be due to the separation of cooperative elements that must work together to induce *Bmp4* transcription in that particular cell type. To test for cooperative elements, the 5’ and 3’ BAC transgenes could be linked together and tested in mouse embryos (Brandt et al. 2008). Alternatively, since the BAC interval tested is only a portion of the gene desert surrounding *Bmp4*, regulatory elements required for *Bmp4* expression in the eye or multiple other structures may be present beyond the interval tested. In support of this, significant noncoding conservation is present in the desert outside of the BAC interval that was tested (data not shown) and would be interesting to test in future studies.

CHAPTER IV

COPY NUMBER ESTIMATION IS SUGGESTIVE OF BAC TRANSGENE INTEGRITY

Introduction

Bacterial artificial chromosomes (BACs) have been used extensively for mouse transgenesis (Heintz 2001) (Giraldo et al. 2003) (Heaney and Bronson 2006). Due to their large insert size, they can often accommodate the complete structure of genes of interest, including long-range *cis*-regulatory elements required for correct tissue-specific or temporal expression. They are also thought to be more resistant to position effects than smaller transgenes (Giraldo and Montoliu 2001) (Gong et al. 2003). For these reasons, they are particularly useful for studying long-range *cis*-regulatory phenomena (Mortlock et al. 2003) (Chandler et al. 2007b) and for experiments where precise transgene expression is critical, such as Cre-recombinase drivers (Lee et al. 2001) (Copeland et al. 2001). In addition, BACs are increasingly used for rescue experiments or overexpression studies. In general, there is little published data that provides detailed documentation for potential correlations between BAC transgene copy number, expression, and structure. More data would be useful regarding the general variation of BAC copy number in transgenic mice and how this variation impacts BAC transgene expression and/or structure. However, the large size of BACs also makes it harder to analyze transgene structure following integration into the genome. Founder animals or their transgenic progeny can provide large

amounts of DNA for Southern blot analysis, although for some developmental studies where transgenic embryos generated by pronuclear injection are analyzed “transiently”, little DNA (e.g. from yolk sacs) is usually available for analysis. PCR-based methods, while limited in scope to analyze large-scale transgene structure, can be useful for estimating transgene copy number. Quantitative PCR (Q-PCR) can also be easily applied to many DNA samples in parallel and provides results faster than traditional Southern blotting, with similar accuracy as we and others have shown (Ballester et al. 2004).

Our laboratory uses BAC transgenes to study long-range *cis*-regulatory elements of the BMP family genes *Gdf6*, *Bmp2* and *Bmp4*. The nature of these experiments depends on verification of BAC transgene structure following transgenesis in mice. Towards this end, we have generated numerous BAC transgenic mice using standard pronuclear injection methods and with several unique BAC transgenes, that were useful for documenting trends in BAC copy number and integrity across independently created transgenic mice. Here, we present a straightforward method for estimating BAC transgene copy number in multiple *Bmp2* and *Bmp4* BAC transgenic lines and embryos using quantitative real-time PCR. In all, we analyzed copy number in 78 transiently generated BAC transgenic embryos or liveborn animals created by pronuclear injection, as well as 317 transgenic mice from 26 separate breeding lines established from liveborn founders. Eleven distinct *Bmp2* and *Bmp4* BAC constructs were used to generate this data. To our knowledge, this is the most extensive analysis to date of copy number in BAC transgenic mice. Our method relies on comparing data

from transgenic samples to a standard curve of calibrator samples that are generated by diluting purified BAC DNA over a range of known concentrations into wild-type mouse genomic DNA. This method is robust, conceptually simple, and amenable to processing large numbers of purified tail DNA samples in parallel. Our data allowed us to confirm stability of several BAC transgenic lines through germline transmission and to correlate copy number with strength of transgene expression. We also observed that transgenic lines carrying multiple BAC copies most likely carry one or more full-length BAC molecules. In general, transgene copy number was fixed in subsequent generations following germline transmission; however, we noticed several examples of striking discrepancies between founder copy number estimates and their transgenic progeny. We also clearly identified several founder animals that each transmitted two independently segregating transgene insertions. Although BACs are extremely useful as transgenic vectors and it is very feasible to create transgenic BAC lines that carry multiple, complete BAC molecules, BAC fragmentation and integration of BACs into separate genomic locations was observed at a frequency of 17% (3/18) and 12% (3/26), respectively. In summary, the monitoring of BAC transgene copy number can add useful information when interpreting BAC transgene expression and confirming stability of integrations through the germline.

Material and Methods

Transgenic Mice

Bacterial artificial chromosome (BAC) vectors were modified using homologous recombination in *E. coli* essentially as described (Mortlock et al. 2003) to contain a *lacZ*:Neo (β-geo) fusion cassette into the *Bmp2* or *Bmp4* transcription unit. Briefly, mouse *Bmp2* BACs RP23-85O11 (239,101 kb) and RP23-409L24 (209,640 kb) were modified as previously described (Chandler et al. 2007b). Mouse *Bmp4* BACs RP23-26C16 (227,097 kb) and RP23-145J23 (227,220 kb) were modified by inserting a GFP-(IRES)-β-geo cassette into the ATG start codon of *Bmp4*. Purified BAC DNA constructs were used for pronuclear injections to generate founder mice and lines as previously described (Chandler et al. 2007b). BAC DNA was prepared and injected as described in Chapter III.

DNA Isolation

DNA was extracted from mice tail biopsies or embryonic yolk sacs by overnight digestion in 500 μL of proteinase K buffer (10 mM Tris-HCl [pH 8.0], 100 mM NaCl, 10 mM EDTA [pH 8.0], 0.5% sodium dodecyl sulfate, 0.25 mg/mL proteinase K) with occasional vortexing. Following digestion, 250 μL of phenol and 250 μL of chloroform was added followed by vigorous vortexing to ensure thorough mixing of phenol:chloroform with the sample. Samples were immediately subjected to microcentrifugation at 16,249 rcf for 4 minutes to allow

separation of the aqueous and organic layers. The aqueous layer was removed with a wide bore pipet tip paying careful attention to avoid the interface. Ethanol precipitation of the aqueous layer was performed and DNA pellets were washed with 70% ethanol followed by resuspension overnight in 200 μL (tail DNA) or 100 μL (yolk sac DNA) of TE [pH 7.4]. Genomic DNA samples were quantified on a UV spectrophotometer at 260 nm and diluted to 10 ng/μL for real-time PCR.

Standard Curve Samples for Real-Time PCR

To create a standard curve of real-time PCR data from known amounts of BAC template, supercoiled BAC DNA was isolated by cesium chloride density centrifugation and quantified via gel electrophoresis, by comparing intensity of restriction-digested BAC DNA bands to lambda DNA/HindIII mass standards as described for the preparation of BAC DNA for pronuclear injections. Then, two-fold dilutions of BAC DNA were spiked into 10 ng/μL genomic DNA (final concentration) that had been isolated from a C57BL6J x DBA2J F1 mouse liver by methods described above and quantified by UV spectrophotometry at 260 nm. This created a series of standard samples such that the ratio of BAC molecules ranged from ~1 to ~ 48 BAC copies per diploid mouse genome. Copy number standards were exposed to at least one freeze-thaw cycle prior to use, since tail and yolk sac DNA samples were also freeze-thawed before analysis.

Real-Time PCR

Custom Taqman® Assays-by-Design were used to generate primer and probe sets for Neo (present in β-geo fusion gene) and the mouse *Jun* gene (control) Applied Biosystems Inc. Assay IDs: 185300786 and Mm00495062_s1. The following primer pairs and probes were used: for Neo assay, forward primer: (5'- ATGACTGGGCACAACAGACAAT-3'); reverse primer: (5'-CGCTGACAGCCGGAACAC-3'); probe: (5'-FAM-CTGCTCTGATGCCGC-3'); for Jun assay, forward primer: (5'- GAGTGCTAGCGGAGTCTTAACC-3'); reverse primer: (5'- CTCCAGACGGCAGTGCTT-3'); probe: (5'-FAM-CTGAGCCCTCCTCCCC-3').

Real-time PCR was performed on a GenAmp9700 thermocycler and plates were scanned using the ABI PRISM® 7900HT sequence detection system. Two microliters (20 ng) of genomic DNA samples or copy number standards were analyzed in a 10 μL reaction volume with two primer-probe sets (Neo, Jun). In addition, no-template controls were included in each experiment. All reactions were performed in duplicate or triplicate.

Copy Number Estimation

Copy number estimates were derived from delta Ct values for standard curve samples. To calculate delta Ct values, the average of duplicate Ct values generated with the Neo probe was subtracted from the average *Jun* Ct value. Using the scatter plot chart function in Microsoft Excel, delta Ct values for each standard were plotted (on the Y axis) against the known copy number of each

standard (on the X axis) using a logarithmic scale. A logarithmic regression trendline and its corresponding equation were then generated to fit the slope. The resulting equation (of the form: y = (slope)ln(x) + y intercept) was used to estimate copy number of samples based on the delta Ct value. To solve for copy number (x), the base of the natural logarithm was raised to the power of X and multiplied by 2 to account for a diploid genome (estimated copy number = $2e^{(\text{deltaCt-y intercept})/\text{slope})}$).

Quantitative Dot Blot Hybridization

Genomic DNA samples were extracted from liver samples isolated from liveborn transgenic mice using standard genomic liver DNA isolation methods described above. Copy number values for the IRES-β-geo cassette were confirmed by dot-blot Southern hybridization using the following method: Copy number estimates were derived from standard curve samples. Standard curve samples were C57BL/6J x DBA/2J F1 hybrid genomic DNA samples spiked with known quantities of pIBG-Ftet plasmid DNA samples diluted to copy number equivalents (1, 2, 4, 8, 16, 32, 64, and 128 copies per diploid genome). 50 μL of standard curve and genomic DNA samples containing 10 μg of total DNA were added to 150 μL of denaturing solution (0.01 M EDTA [pH8.0], 0.53 N NaOH). Samples were incubated at 95°C for 5 min., and then placed on wet-ice for 2 min. A Zeta-Probe GT membrane (Bio-Rad) was briefly washed twice with H_2O then once with 0.4 N NaOH for 5 minutes. The pre-washed membrane was placed on a 96-well Minifold Vacuum Filtration Manifold apparatus (Schleicher and Schuell),

and the apparatus was assembled according to the manufacturers' instructions. Denatured DNA samples (200 μL total volume for each) were loaded onto the vacuum manifold and incubated for 30 min. at room temperature. Following incubation, samples were vacuum filtered for 5 minutes until all of the samples had passed through the membrane. The membrane was then neutralized with 0.2 M Tris-HCl [pH 7.5], 2x SSC (1x SSC is 0.15 M NaCl, 0.015 M sodium citrate) for 10 min and baked for 30 min at 80°C. Control genomic probe used in hybridizations was generated by PCR amplification of mouse genomic DNA using primers specific to the 3' UTR of mouse *Gdf6* (Forward primer 5'-AAGCATGGAAAGAGGATGAAAGGG-3', Reverse primer 5'-CGACCTCCAGTAACTTTAGTGTTGTCA-3') and subsequent cloning into pCRII-TOPO (Invitrogen) followed by restriction enzyme digestion with NotI and SpeI to isolate a ~937 bp fragment. The transgene-specific probe used in hybridizations was generated from a 4.7 kb XbaI fragment containing the IRES-β-geo cassette from pIBG-Ftet (described above). Both control and transgene-specific probes were labeled using Ready-to-Go labeling beads (Amersham) and [α-^{32}P]dCTP (Amersham). For control probe hybridizations, the membrane was washed twice with sterile H_2O and hybridized for 3 h with Rapid-Hyb buffer (Amersham) according to the manufacturers' instructions. The membrane was then exposed to a Kodak phosphor-imaging screen for 5 days and imaged using a Pharos FX imaging system and Quantity One® software (Bio-Rad). Immediately following exposure, the membrane was placed in strip buffer (10 mM Tris-HCL [pH 7.6], 1 mM EDTA [pH 8.0], 1% sodium dodecyl sulfate) and boiled for 10 minutes to

remove bound probe. The membrane was neutralized and baked (described above), then processed for transgene-specific probe hybridization. Again, the membrane was washed twice with sterile H_2O and hybridized for 3 h with Rapid-Hyb buffer (Amersham) according to the manufacturers' instructions. The membrane was then exposed to a Kodak phosphor-imaging screen for 16 hours and imaged using a Pharos FX imaging system and software (Bio-Rad). For both control probe and transgene-specific hybridizations, triplicate standard curve and genomic DNA samples were measured using Biorad imaging Quantity One® software. Copy number estimates were derived from standard curve samples. Control genomic probe hybridizations were used to calibrate total input DNA.

Preparation of Agarose-Embedded High Molecular Weight DNA from BAC Transgenic Embryos

Mouse embryonic fibroblasts (MEFs) used for the generation of agarose-embedded high molecular weight DNA were isolated and cultured from e13.5 embryos generated by crossing BAC transgenic males with wild-type Crl:CD1 (ICR) females as described previously (Chandler et al. 2007b). Cells were embedded in 0.75% low melting point agarose (Invitrogen) using agarose plug molds (Bio-Rad) prior to restriction enzyme digestion and pulsed field gel electrophoresis.

Southern Analysis of High Molecular Weight Transgenic DNA

Agarose plugs (isolated as described above) were washed twice in 50 mL of TEX buffer (10 mM Tris-HCl [pH 8.0], 1 mM EDTA [pH 8.0], 0.01% Triton

X-100) and once in 50 mL of TE pH 8.0 (10 mM Tris-HCl [pH8.0], 1 mM EDTA [pH8.0]) at room temperature with agitation. Plugs were then transferred to 2 mL screw cap tubes (1 plug per tube) and equilibrated for 30 min at room temperature in 2 mL of 1x restriction enzyme buffer (NEB) containing 10 mM spermidine trihydrochloride with agitation. Once equilibrated, the solution was replaced with 800 μl of 1x restriction enzyme buffer containing 10 mM spermidine trihydrochloride and 200 units of MluI (NEB) and incubated for 6-8 h at 4°C on a three-dimensional rotator (Lab-line) to allow the enzyme to infiltrate the agarose plug. After 6-8 h, the tubes were transferred to a 37°C incubator and incubated for 12-16 h with agitation. Plugs were then washed twice in 2 mL TEX buffer at 4°C for 30 min each with agitation and equilibrated at room temperature in 2 mL of 0.5x Tris-Borate-EDTA gel electrophoresis buffer. Restriction digested high molecular weight DNA was then resolved by pulsed-field gel electrophoresis in 1% agarose/0.5x Tris-Borate-EDTA buffer for 20 h at 15°C (6 V/cm, 6-80 s switch time). DNA fragments were depurinated with 0.25 M HCl for 20 minutes and transferred by alkaline capillary transfer onto a Zeta-Probe GT membrane (Bio-Rad) with 0.4 N NaOH for 24 h. The membrane was neutralized with 0.2 M Tris-HCl [pH 7.5], 2x SSC (1x SSC is 0.15 M NaCl, 0.015 M sodium citrate) for 10 min and baked for 30 min at 80°C. Probe used in hybridizations was generated from a 4.7 kb XbaI fragment containing the IRES-β-geo cassette from the pIBG-Ftet (described above), and was labeled using Ready-to-Go labeling beads (Amersham) and [α-^{32}P]dCTP (Amersham). Membranes were washed twice with

sterile H_2O and hybridized for 3 h with Rapid-Hyb buffer (Amersham) according to the manufacturers' instructions. Membranes were exposed to a Kodak phosphor-imaging screen for 16 h and imaged using a Pharos FX imaging system and Quantity One® software (Bio-Rad).

Expression Analysis of Transgenic Mice

To generate embryos for XGal staining, transgenic male mice were crossed to Crl:CD1(ICR) female mice to obtain timed pregnancies. Pregnant mice were sacrificed by CO_2 inhalation and their embryos were harvested for XGal staining to detect *lacZ* expression. In brief, embryos were dissected into 1x phosphate buffered saline (PBS) on ice then fixed with 10% neutral buffered formalin at 4°C with agitation. Embryos older than 14.5 days post coitus (dpc) were bisected to allow for reagent penetration after fixation. Next, embryos were processed for XGal staining essentially as described (DiLeone et al. 1998) with two minor changes: 1) 0.6 mg/mL XGal was used and 2) embryos were stained overnight at room temperature with agitation.

Polymorphic Marker Analysis of *Bmp4* BACs

Polymorphic marker analysis was performed along the length of each BAC using primers designed to amplify simple tandem repeats (STRs) that are polymorphic between C57BL/6J and DBA/2J strains. Since the BACs are derived from the C57BL/6J strain and the transgenic founders were (C57BL/6J x DBA/2J) F_2 hybrids, we could identify some transgenic founders that were

fortuitously homozygous for DBA/2J alleles at the *Bmp4* locus by genotyping STR markers flanking the *Bmp4* region but outside the BACs. For some lines, animals were backcrossed to DBA/2J mice to generate the required DBA/2J homozygotes. Once we identified mice that were DBA/2J homozygotes for flanking STRs, markers designed to assay the C57BL/6J derived BAC were used to interrogate the integrity of the transgene. The flanking markers (centromeric: D14Mit212, D14Mit56; telomeric: D14Mit141, D14Mit60) were identified using the JAX MGI database (www.informatics.jax.org). Internal STRs were identified within the BAC insert sequences obtained from the UCSC genome browser using custom software and a subset of polymorphic STRs were identified by comparing PCR products from C57BL6/J and DBA/2J DNA samples for length differences. Primers were designed to amplify polymorphic STRs in both the flanking sequence and internal sequence (TABLE 4.1). Simple tandem repeat (STR) markers internal to the BMP BACs used in this study were identified by custom algorithms (K.M. Bradley and J.R. Smith) set to screen for tandem repeats with a sequence ranges of 2-6 and a minimum number of ten repeats present. Primers flanking these repeats were designed to amplify PCR products of less than 275 bp that were then screened for length variations between C57BL/6J and DBA/2J mouse strains by electrophoresis on non-denaturing polyacrylamide gels (Deal et al. 2006). STR PCR products that displayed detectable variations in length were used to evaluate the presence of C57BL/6J *Bmp4* BAC sequences in transgenic animals.

Table 4.1 Primer sequences used for polymorphic marker analysis

Target	Primer Name	Forward Primer (5'-3')	Reverse Primer (5'-3')
5' Flank	D14Mit60	AGGCTGCCCATAAAAGGG	GTTTGTGCTAATGTTCTCATCTGG
	D14Mit141	CCAGCATTCCGAAGTCATTT	AGGGAAAGAAGACAGCACGA
3' Flank	D14Mit56	TGGCAAAGTTTTTTTTTTCCC	TCTGGGTAGAACTGTAATAGCACA
	D14Mit212	AACATGTGCACTGGAACAATG	TCATTTATCAATTTACTTTGGTGAGG
5' Internal	C8	AGATACTCTAGCTGGGGC	GCTGTGCACGATTGTTA
	E8	CAATCCCCAGCTCAAAAC	GGAAGGTAGCTTTCCATC
	A9	CCATTACCCAGTCATGAC	AAGTAAGCCATTGCCTC
	C9	ACAGCTCACAGTTTGAGC	AGGTGTGTGAACTTGAAC
	E9	CAGGGTATCAACAGGAAC	CATGTAGCTAAATCTTGCC
	G9	CTGATGCTTCAAGTTACAC	CAAAGTTCCTTCTGAGGT
	C11	ACAGCAAAGGTCTCAGAC	GGGGTTTCAGCTCAGTAA
	E11	CTTGGCCCATTTCTTTAC	AGTGTGCATGTATGTGCA
Overlap	G6	TAGCTCCAGCACTTTGG	CAGAAGACAAGGTCATTCT
	A7	TGAGGGACAAGCAGTAGT	TTACAGCCTCCAATCCA
3' Internal	A1	CATGTGAGATCTAGGCTC	CAGGCTGATAGTTCCTAAG
	E1	AGAACACTGGCTGCTCTT	GCTTGCTTGTATGTCATG
	G1	AGCAACAGCATCTTCTGG	GATGGCACTCATGCACTC
	A2	GGTATCTGCATACACATGC	CCAAACAGTGACCACTTT
	A3	GTTGAGATTCTATTGTCCC	GTCTCAGAAATGTTGAGAAG
	E3	GTCTCAGAAATGTTGAGAAG	ACGGAATTATTGGTAGCC
	G3	AGAAACCCATAGGGCTG	AGATGAGTGTTCCCCTTA
	1	GTACGTGTTTCTCAGACTC	CTGATTTGAGTTTCCTATC
	11	GTCCTCCATTTCTTCTT	GGCTCGATACAGAAAGCT
	E5	TTTCAACCATGAGTGGT	CATACACACTTGCATGCT
	A6	GGCATGGCATACACACTA	CGCCTGGTAGGATGTACT

Results

Validation of Method for Estimating BAC Copy Number by Real-Time PCR

Accurate estimation of copy number in unknown samples relies on the use of accurate standards and an effective linear range of detection. Our estimations were based on comparisons to a standard curve. We reasoned that by comparing samples of unknown copy number to a range of DNA copy number standards, each made with the same amount of mouse genomic DNA spiked with varying amounts of purified BAC DNA, we would be able to extrapolate copy number estimates in a manner that would help control for differences in amplification efficiency between PCR assays. This curve was comprised of real-time PCR data from standard template samples, which contained two-fold dilutions of known quantities of BAC DNA in the same concentration of mouse genomic DNA. The BAC dilutions were designed to represent a range of approximately 1-48 copies of BAC molecules per diploid genome. For each standard, Ct values were generated using both an assay specific to the transgene (Neo) and an assay for a nontransgenic control gene (Jun). Amplification plots of the BAC copy number standards showed that all standards amplified similarly for the internal control (Jun) and that standards showed a stepwise one cycle difference in Neo assay profiles, as expected (FIGURE 4.1a). Delta Ct values were plotted to generate a standard curve. Standard curves were highly similar in independent experiments with coefficients of determination (R^2) close to one, indicating the dilutions were made accurately, gave consistent results, and could be used to generate curve equations for estimating copy

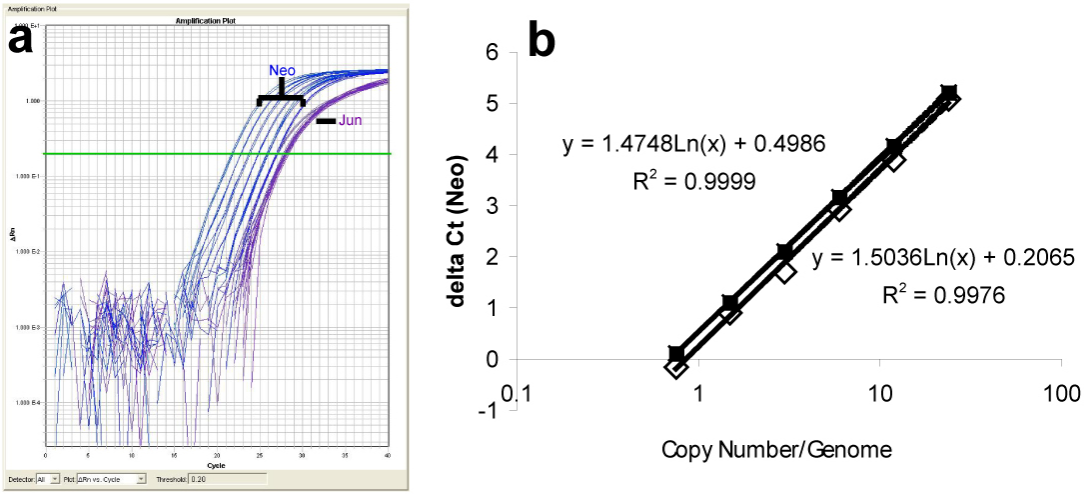

Figure 4.1. BAC DNA copy number standards generate reproducible curves in real-time PCR. (a) Amplification plot depicting the Neo (blue) and Jun (purple) results for the copy number standards. As expected, each standard is approximately one cycle apart for the Neo assay and amplification plots are similar for all standards for the Jun assay. (b) Copy number standards were used to generate standard curves in real-time PCR using Neo and Jun primer/ probe sets on two independent days (filled boxes = Day 1, empty boxes = Day 2). Replicate experiments indicate the copy number standards provide highly reproducible standard curves (R^2 = 0.9999, R^2 = 0.9976) for estimating copy number of experimental samples.

number from actual transgenic animals (FIGURE 4.1b). Therefore, the copy number standards provided a method for generating robust standard curves.

We then tested dilutions of experimental samples to examine whether copy number estimates varied substantially depending on the amount of input genomic DNA used as template. Two-fold dilutions of genomic DNA were prepared from two independent BAC transgenic animals from different lines, and used as templates for real-time PCR (FIGURE 4.2). This indicated that copy number estimates varied little over a linear range of input DNA from 4 – 32 ng. Therefore, 20 ng of genomic DNA (based on spectrophotometer readings) was used as input DNA for PCR, since copy number estimates varied little with this amount of input DNA. Likewise, 20 ng of mouse DNA was used in the standard samples. We performed our initial real-time PCR experiments in triplicate and found very close data points across replicates. Therefore, we reasoned that if there were no significant difference between performing the experiment in duplicate reactions versus triplicate reactions, we would perform the remainder of our experiments in duplicate. To verify there was no significant difference between the experiments performed in duplicate versus triplicate, we performed a paired T-test on an experiment encompassing 16 individual mice from several lines of varying copy numbers for which triplicate-averaged results were compared to duplicates. We found results generated by real-time PCR performed in duplicate were not significantly different from results generated by triplicate reactions ($p>0.14$). Therefore, we performed the bulk of our real-time PCR experiments in duplicate unless otherwise noted.

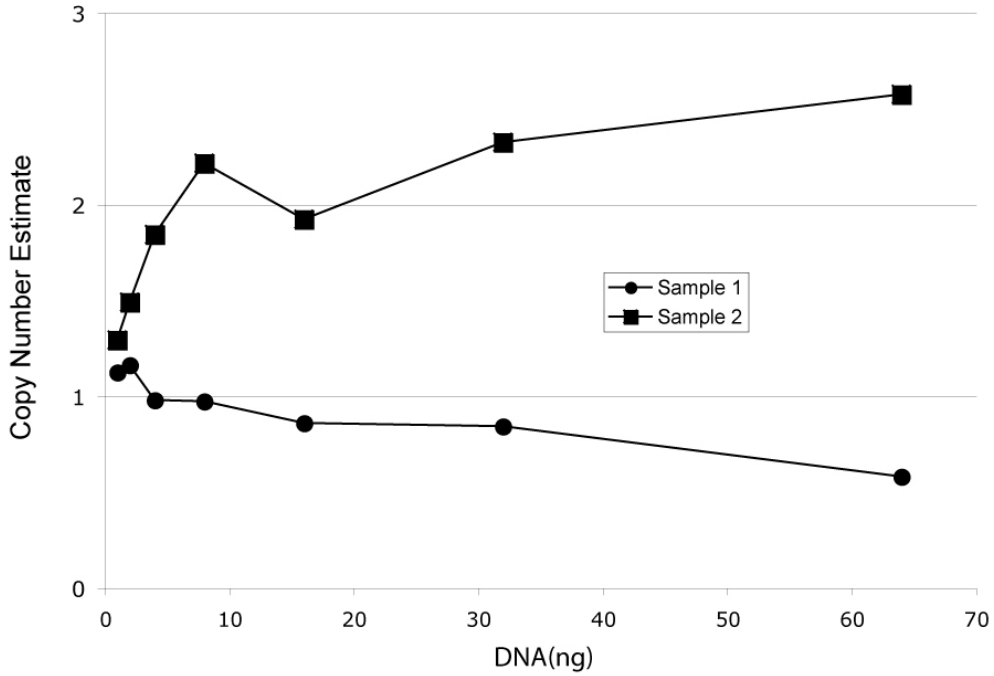

Figure 4.2. DNA concentration has little impact on copy number estimates over a wide range of input DNA. DNA samples from two transgenic mice with copy number estimates of 2 (filled box) and 1(filled circle) were each used to create a 2-fold dilution series of DNA templates, such that 1-64 ng DNA (total input) from each mouse were subjected to real-time PCR. The amount of template DNA versus copy number estimations indicate copy number estimations vary little as input DNA ranged from 4-32 ng.

By fitting delta Ct values from experimental samples to the standard curve equation, we estimated BAC copy number for a total of 78 transgenic founders (embryos and liveborn mice), and for 317 mice from 26 independent transgenic lines that were established from breeding some of the liveborn founders. We first looked for evidence to confirm consistency of our estimation method across samples. When estimating copy number for multiple transgenic littermates, we found copy number values were generally consistent among littermates within a line. For example, DNA samples isolated from yolk sacs or tails from littermates of independent lines produced similar copy number estimates, with minimal variability between DNA samples within a litter (FIGURE 4.3). Lines with the highest copy numbers had sample estimates that fell outside the linear range of our standard curve; not surprisingly, these showed a wider range of copy number estimates (FIGURE 4.3b).

Conventional methods for estimating copy number in transgenic lines include quantitative dot blot hybridization or Southern blot analysis. To further validate the copy number estimates from real-time PCR data, we performed quantitative dot blot hybridization on genomic DNA samples purified from livers from a limited number of mice from six *Bmp4* BAC lines (FIGURE 4.4) and compared estimates based on dot blots to those generated by real-time PCR using the same individual liver DNA samples (TABLE 4.2). To control for differences in amount of input DNA, a control genomic probe was utilized (FIGURE 4.4b). Both the dot blot and real-time PCR analysis of liver DNA samples were performed in triplicate. Statistical analysis revealed no significant

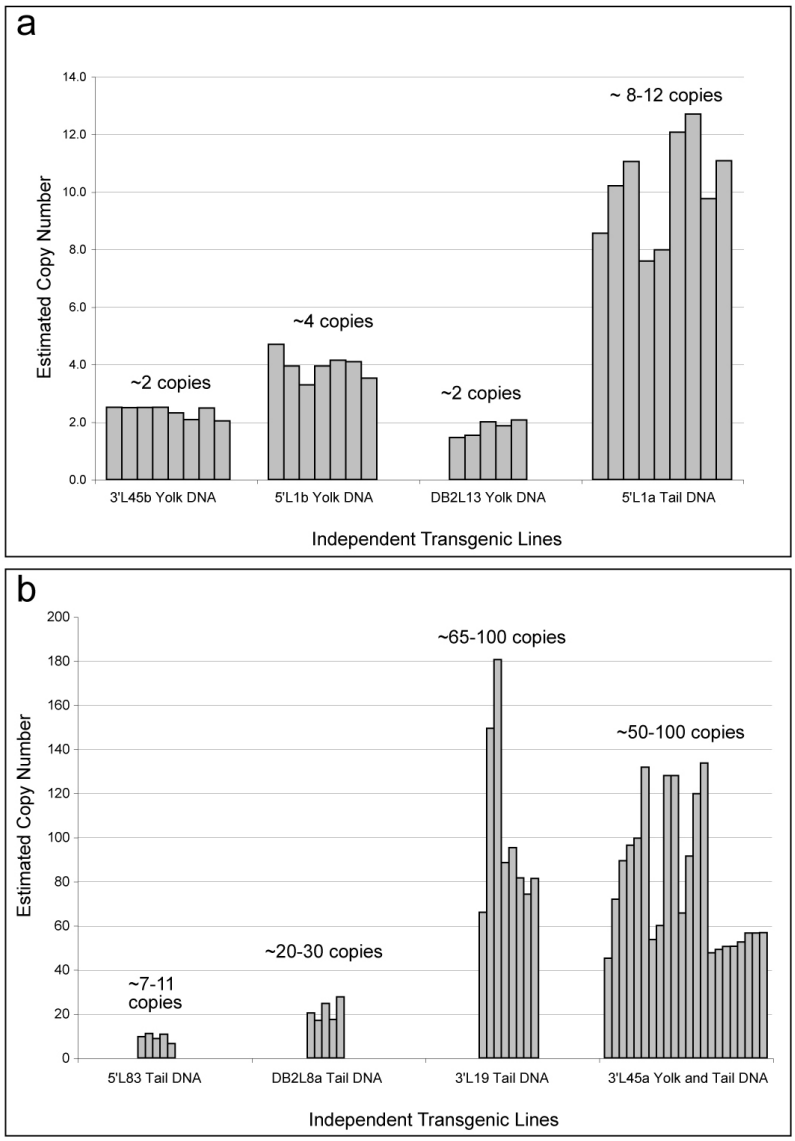

Figure 4.3. Copy number estimates are consistent within independent transgenic lines. Shown are copy number estimates from individual mice, as determined from yolk or tail DNA samples of multiple progeny from eight independent *Bmp4* BAC transgenic lines.

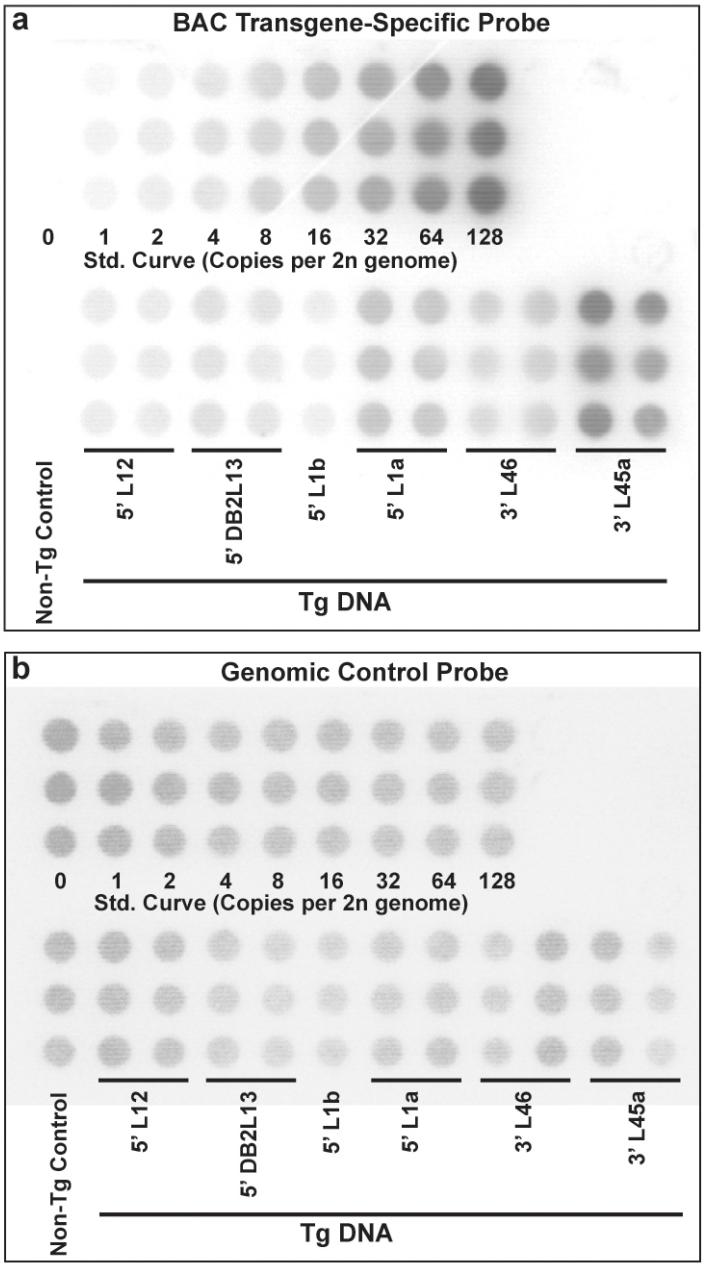

Figure 4.4. Copy number estimation by dot blot hybridization. (a) Dot blot hybridized with a BAC transgene-specific probe (IRES-βgeo). (b) The same dot blot stripped and reprobed with a genomic control probe (mouse *Gdf6* 3'UTR fragment) to account for differences in input DNA. Standard curve (0-128 copies per diploid genome) and genomic DNA samples were assayed in triplicate. Copy number estimates for genomic DNA samples were derived by comparing the ratio of dot intensities for transgene-specific and control probe hybridizations in standard curve samples spiked with known quantities of pIBGFTet plasmid (see methods).

Table 4.2. Comparison of copy number estimates generated by dot blot analysis versus real-time PCR on Bmp4 BAC transgenic mouse liver DNA samples from individual mice.

Line	Mouse	Dot blot (Liver)	Avg.	RT-PCR (Liver)	Avg.	RT-PCR avg. for line
3' BAC L46	1	14	14	14	12	10
	2	13		10		
3' BAC L45a	1	76	77	75	83	76
	2	78		90		
3' BAC L13	1	7	8	4	5	4
	2	8		5		
5' BAC L12	1	2	3	2	3	2
	2	3		3		
5' BAC L1a	1	22	20	15	16	11
	2	17		16		
5' BAC L1b	1	5	N/A	4	N/A	4

difference between copy number values estimated by traditional dot blot analysis and values estimated by our real-time PCR methods (paired T-test, p=0.74). The average real-time PCR copy number estimates generated from tail DNA samples of multiple transgenic mice within a line (TABLE 4.2, far right column) was close to the copy number averages estimated by dot blot from two mice. Therefore, the real-time PCR estimation method seemed suitable for application to many DNA samples from tail snips or embryonic yolk sacs.

Since real-time PCR results can be skewed by contaminating materials that adversely affect amplification, we examined our data set for evidence of consistent DNA quality. To do this, for 26 transgenic lines we specifically

computed the average copy estimate for all transgenic animals (317 mice from 26 breeding lines), and we counted all animals having individual estimates that were within two-fold of the initial average for the line. (For this, we only considered lines that were clearly segregating only single sites of transgene integration.) All estimates were based on duplicate Neo and Jun PCR reactions. We observed 291 of 317 DNA samples (47 yolk sac and 244 tail DNA samples) tested gave copy number estimates that were within this range. A limited number of samples (7 yolk sac and 19 tail DNA samples) gave estimates that were either two-fold greater or less than the initial average, and were considered poor quality DNA isolations. These estimates may have been skewed by issues relating to impure DNA template; previous reports state that materials in mouse tail tissue or traces of phenol can inhibit PCR (Burkhart et al. 2002). The majority of both yolk sac and tail DNA samples were within two-fold of their line average (291/317). Nevertheless, this was a limited problem that was easily overcome by examining multiple animals for a given transgenic line (e.g. FIGURE 4.3). While we also reasoned that our copy number estimates based on single samples should be interpreted with caution, we concluded this method would still be useful for analyzing copy number trends across many founder animals.

Distribution of Copy Number Across Breeding Lines and Founders

In our laboratory, we have generated a number of BAC transgenic embryos and breeding lines as part of our efforts to study regulation of the *Bmp2* and *Bmp4* genes. This involved eleven unique BAC constructs (six *Bmp2* and

five *Bmp4* BACs). While much of our data was based on progeny resulting from germline transmission of transgenes, some of our copy number data was based on DNA from mid-gestation transient transgenic embryos or liveborn mice generated directly from pronuclear injection. We refer to these as "transient transgenic" mice. Upon analyzing copy number estimates for 78 transient transgenic embryos or founder mice, and for 26 breeding transgenic lines established from some of the founders, we were able to compile this data and create a distribution of copy number values for each independent founder or average values for each breeding line (FIGURE 4.5). For this analysis, we recalculated copy number estimates for all transgenic lines after excluding the poor quality DNA samples as defined above. As described below, we found several cases where two independently segregating insertions were clearly derived from one founder animal. In these cases, the separate insertions were considered as separate lines. For each line, samples from at least three mice were used to generate the average copy number value (avg. number of mice used for each line = 13).

Investigation of copy number in both transient transgenic embryos or mice and breeding lines allowed us to compare all BAC transgene integration events (FIGURE 4.5). This distribution clearly shows that the majority of transient transgenic embryos, liveborn founder mice, and breeding transgenic lines contained one or more transgene copies per genome (FIGURE 4.5). Not unexpectedly, real-time PCR analysis suggested that every breeding line contained one or more transgene copies. Approximately 22% of transient

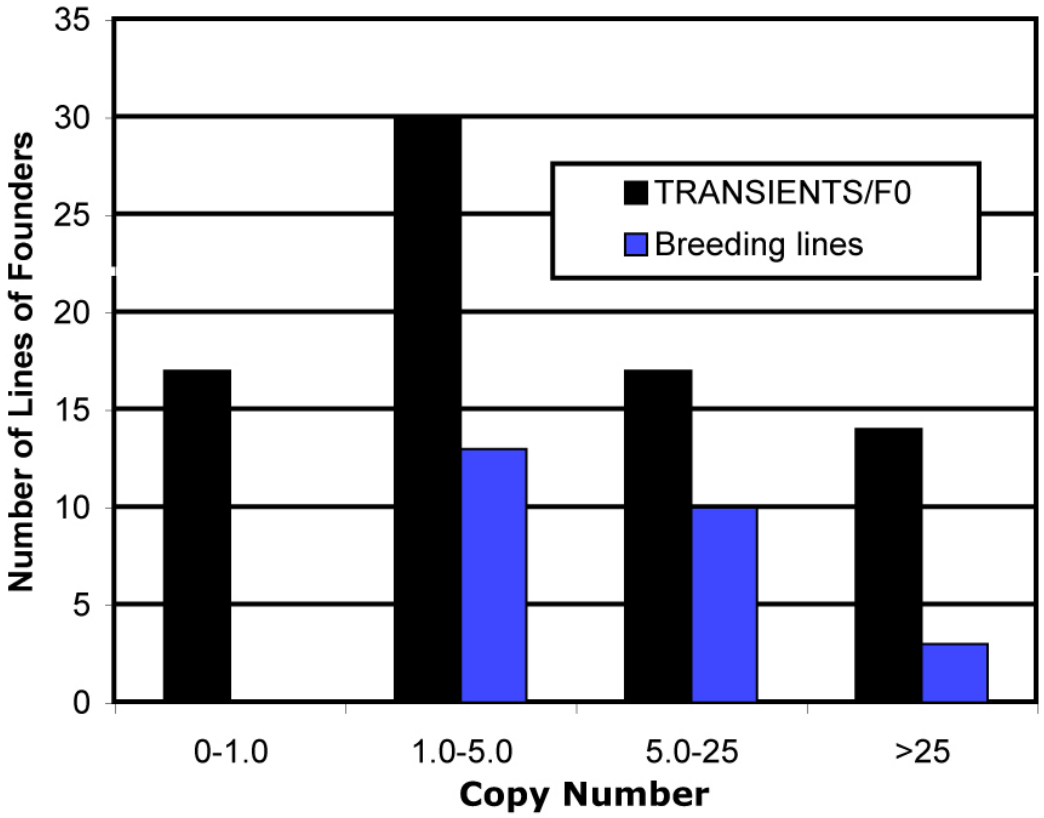

Figure 4.5. The distribution of variation in copy number across stably breeding lines and transiently generated founder embryos or liveborn founder mice. While the majority of integration events contain 1-25 copies, all animals with estimates of fewer than one copy per genome were founder animals, suggesting somatic mosaicism.

transgenic or founder animals (17/78) were estimated to carry less than one transgene copy per genome, suggesting many of these animals were probably genetically mosaic for the transgene. However, such founders could often produce transgenic offspring, in some cases with rather high transgene copy numbers (not shown), suggesting that somatic mosaicism in the founder can preclude the ability to predict copy number estimates in their offspring before actually breeding the founder. Although most published reports suggest BAC transgenes integrate as low copy concatamers (Jaenisch 1988) (Giraldo and Montoliu 2001) (Heaney and Bronson 2006), 18% of transient transgenic embryos or founders (14/78) had copy number estimates greater than ~25 BAC copies per genome. After breeding a subset of founders, 12% (3/26) of established lines also had more than ~25 copies, although founder estimates did not always predict the high copy numbers in offspring (see below).

Analysis of Copy Number in Successive Generations

Although most transgenic mice made via pronuclear injection have transgene DNA inserted at a single genomic location, integration into two separate, unlinked locations can occur (FIGURE 4.6). As expected, most of our BAC founder animals (N = 20 of 23 bred founders) generated close to 50% transgenic and 50% non-transgenic progeny and copy estimates were similar among transgenic littermates, consistent with there being a single, stable site of BAC transgene integration in the founder. Our BAC lines are designed to drive *lacZ* expression during mouse development as a convenient reporter for *Bmp2* or

Bmp4 expression. Occasionally, when transgenic founders were bred to generate liveborn progeny and/or embryos for XGal staining, we noticed that roughly 75% of progeny were transgene-positive and that there were obviously two different levels of XGal staining intensity (i.e. “strong” vs. “weak”). This was observed for three of the 23 founder animals, suggesting each founder could transmit at least two distinct, unlinked transgene insertions. Pedigree analysis of Bmp4 5’ BAC line L1 across two generations confirmed evidence for two independent integration events that segregated independently (FIGURE 4.6; for clarity, the non-transgenic littermates are not shown; however, 19 of 25 weanlings from two litters were transgene-positive). Interestingly, one integration in this line has approximately ten BAC copies whereas the other integration has four copies. *lacZ* expression analysis confirmed that embryos generated from stud males containing the “high copy” integration have more robust expression as compared to embryos carrying the “low copy” integration (FIGURE 4.6a). In addition, the copy number estimate obtained from tail DNA of the pedigree founder was close to two, whereas copy number estimates of multiple progeny strongly suggest the founder actually carried two integrations that each had more than two copies (copy number estimates from the founder female were determined from two independent tails snips to confirm these results; estimates were 1.8 and 2.6 copies, respectively). Although copy number estimates for the founder and F1 progeny were different, estimates for successive generations were stable (FIGURE 4.6a and data not shown). Similar to *Bmp4* 5’ BAC L1, pedigree analysis of *Bmp4* 3’ BAC line L45 revealed two independent integration

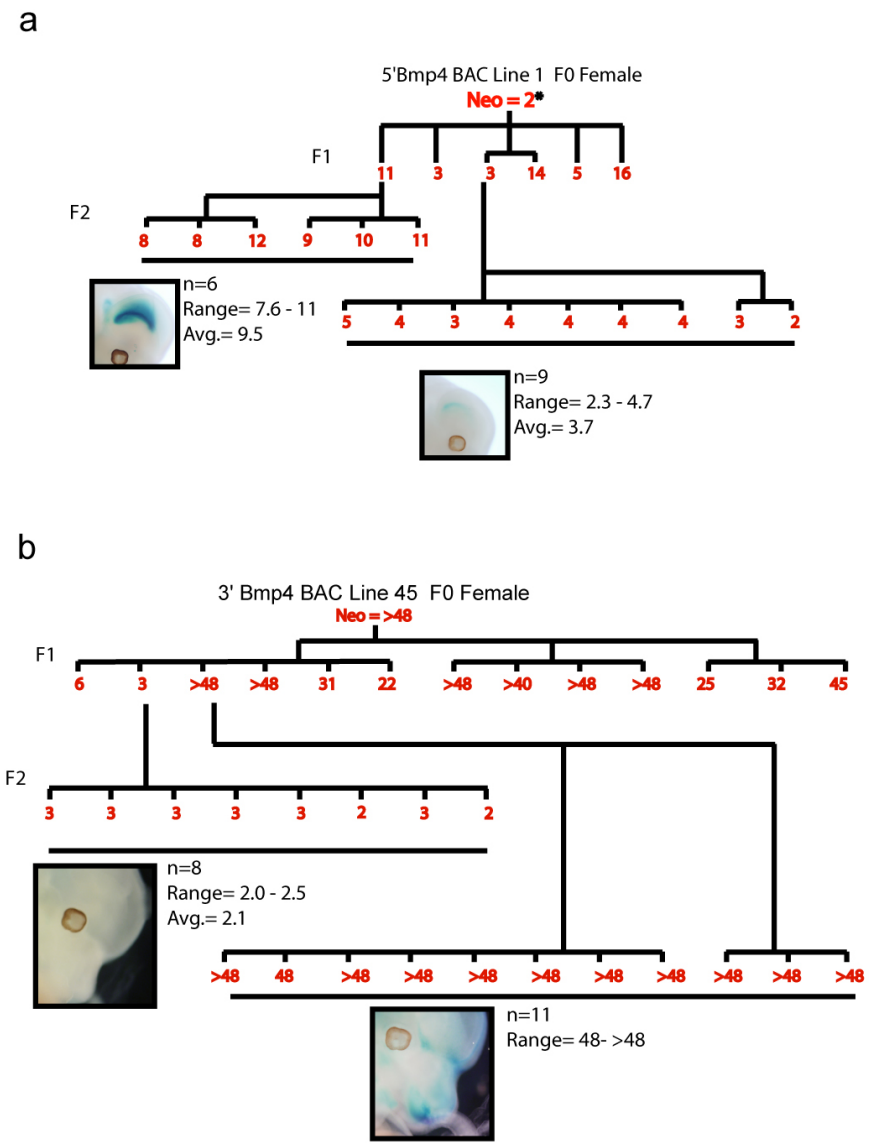

Figure 4.6. Pedigree analysis of mice generated from two independent founder mice reveals that in both cases, BAC transgenes have inserted in two separate, segregating locations in the genome as demonstrated by number estimates. In both cases this was supported by ~75% rate of transgenesis in F1 progeny (non-transgenic littermates not shown). Copy number estimates for individuals are shown in red. Inset images show representative XGal stained e12.5 embryos characteristic of each independent integration event. (a) 5' *Bmp4* BAC Line 1 founder female generated mice with "high" (Avg. = 9.5) and "low" (Avg. = 3.7) copy number estimates that segregate independently. For the founder, copy number estimates were based on the average of two independent tail biopsies/DNA preps (*). (b) 3' *Bmp4* BAC Line 45 founder female generated F1 progeny with "high" (Avg. = >48) and "low" (Avg. = 2.1) copy number estimates that segregate independently.

events (FIGURE 4.6b). In this case, copy number estimates for the founder female were similar to copy number estimates of the “high copy” integration event. A third founder also transmitted two segregating transgene insertions (not shown). Although these three founders each demonstrated multiple integration events, the majority of breeding founders analyzed (20/23 founders) had no evidence for multiple integrations.

Correlation Between Increased Copy Number and Increased Expression

It has been generally observed that for large transgenes, the correlation between strength of expression and copy number is more consistent than for small constructs. However, silencing of gene expression has been reported for large transgenes when present in “high” copy numbers (8-14 copies) (Li et al. 2000). To ensure staining differences were not due to varying protocols, embryos from different lines were stained for the same amount of time at the same temperature. We found that strength of XGal staining in *Bmp2* or *Bmp4* BAC transgenics correlated qualitatively with increased transgene copy number, as shown in Figure 4.7, although since we did not measure expression quantitatively, we cannot determine if expression rigidly correlates to copy number in high copy lines. As previously published, 3’ *Bmp2 lacZ*-BAC transgenic embryos display a subset of endogenous *Bmp2* expression patterns such as whisker hair shaft, ventral footpads, osteoblast progenitors (bone), intervertebral discs, kidney, pelage hair follicle placodes, midbrain, and interdigital mesenchyme (Chandler et al. 2007b). Deletion of a 40 kb segment of

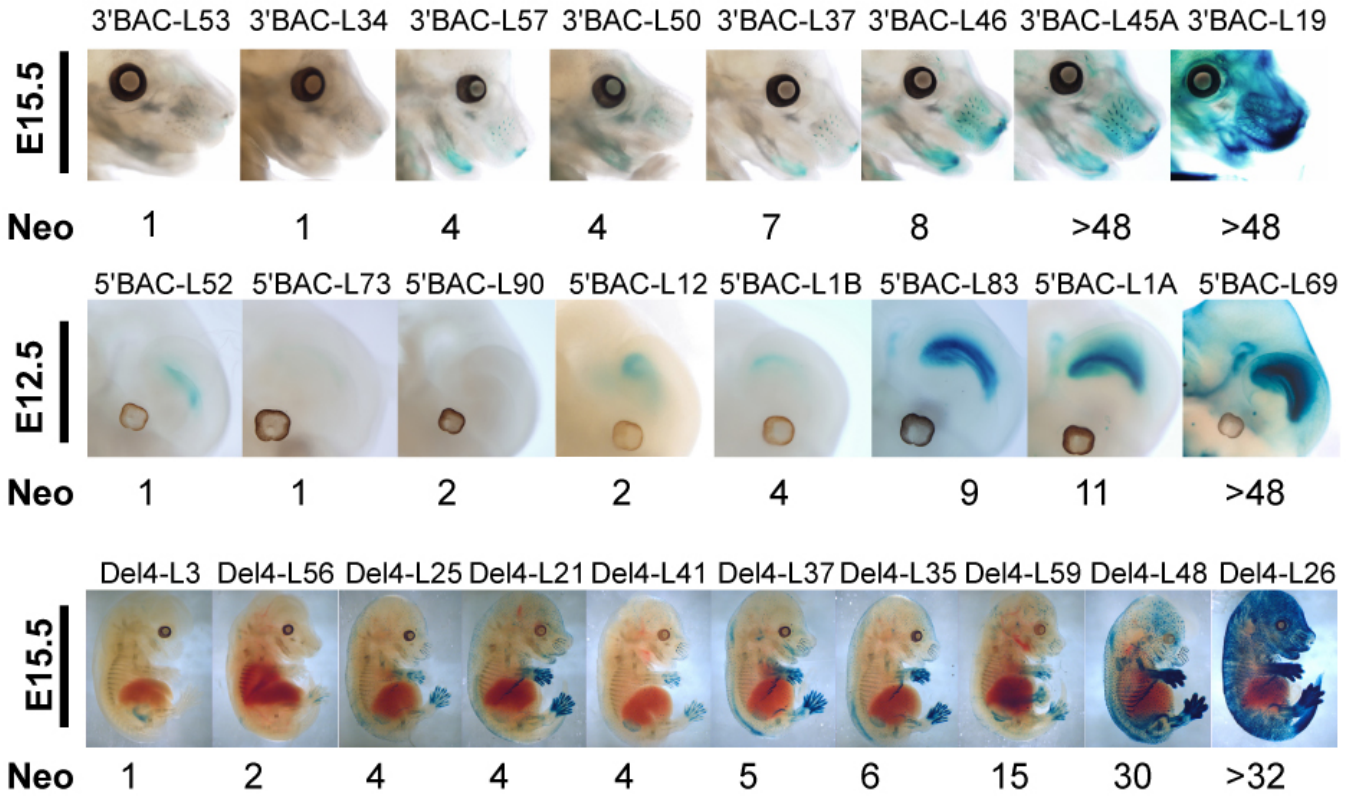

Figure 4.7. XGal stained embryos generated from three distinct BAC transgene constructs suggest that increasing BAC transgene copy numbers correlate with increased transgene expression. Each row of images represents embryos from a separate BAC construct, arranged by increasing BAC copy number estimates. Top row: e15.5 embryos from independent *Bmp4* 3'BAC lines. Middle row: e12.5 embryos from independent *Bmp4* 5'BAC lines. Bottom row: e15.5 *Bmp2* deletion-BAC transgenic founder embryos.

3' *Bmp2 lacZ*-BAC (Del4) results in the tissue-specific loss of BAC-directed *lacZ* expression in the intervertebral discs and the midbrain (Chandler et al. 2007b). *Bmp2* Del4 transient transgenic embryos show little to no *lacZ* expression when copy number is low (FIGURE 4.7, bottom panel). As copy number increases, so does the apparent strength of *lacZ* expression. Of note, close inspection of transient transgenic embryo Del4-L41 showed a mosaic pattern of staining (not shown), thus its overall staining strength appears to be relatively weaker than other embryos of similar copy number (FIGURE 4.7, bottom panel). Although expression is decreased in embryos with modest copy number estimates, staining patterns were very similar to those in embryos with very high copy number. For example, *Bmp2-lacZ* BAC embryo Del4-L25 has expression in limb bones and digits as seen in *Bmp2-lacZ* BAC embryo Del4-L26 albeit at a significantly reduced level (FIGURE 4.7, bottom). This transgene also drives expression in hair follicles, such that the exterior of the embryo appears strongly stained at higher copy numbers (e.g. L48 and L26 embryos) (Chandler et al. 2007b); lower copy number embryos had fainter, distinct expression in hair follicles (not shown). *Bmp4-lacZ* BAC transgenes direct expression faithfully in several tissues where *Bmp4* is endogenously expressed (K. Chandler et al, *manuscript in preparation*). For example, 3' *Bmp4 lacZ*-BAC transgenes direct expression in the craniofacial mesenchyme and whisker hair shafts (FIGURE 4.7, top panel). Both structures are documented sites of *Bmp4* expression (Jones et al. 1991) (Carninci et al. 2005) (Bitgood and McMahon 1995). Likewise, 5' *Bmp4 lacZ*-BAC transgenes direct expression in the developing forebrain, choroid

plexus and whisker primordia where *Bmp4* is known to be expressed (FIGURE 4.7, middle panel) (Furuta et al. 1997) (Bitgood and McMahon 1995). Thus, analysis of *lacZ* expression in embryos generated from these three transgene constructs clearly showed increasingly robust *lacZ* expression as copy number increased, with no evidence of strong silencing effects with higher copy numbers. In contrast, lines with few transgene copies exhibited markedly reduced or completely absent *lacZ* expression. Similar trends were observed with no exceptions in a total of 26 breeding lines, involving the constructs in Figure 4.7 and in eight additional BAC constructs (data not shown).

Analysis of BAC Transgene Integrity

We investigated the possibility that internal deletions within the transgene might have occurred in lines that had minimal or absent expression. Transgenes that are introduced by pronuclear injection typically integrate into the genome as tandem concatamers by a mechanism involving homologous recombination between circularly permuted molecules (Bishop and Smith 1989) (Bishop 1996) (Hamada et al. 1993). Although large molecules can be prone to breakage before integration (Bishop and Smith 1989), leading to insertion of fragmented transgenes, it has been reported that multiple-copy BAC insertions usually have at least one full-length monomer (Gong et al. 2003) (Chandler et al. 2007b); however, BAC transgene integrity and copy number are not always compared in published studies. One approach to monitor integrity of large transgenes is polymorphic marker analysis (Deal et al. 2006). We used this approach to

analyze the presence of polymorphic markers across the length of transgenes in 18 *Bmp4* BAC lines (FIGURE 4.8). This confirmed that multiple-copy BAC transgene integrations most often contain all segments of the transgene that were assayed, suggesting these lines may have at least one intact copy of the BAC transgene. Twenty simple tandem repeat (STR) polymorphisms with an average distance of 19 kb between each polymorphism were assessed within the *Bmp4* BAC transgenes. The majority of lines (15/18) were shown to harbor all transgene-specific polymorphisms, suggesting integration of complete BAC molecules (FIGURE 4.8). However, three lines lacked transgene-specific polymorphisms across one portion of the BAC, suggesting these transgenes integrated into the mouse genome as partial fragments. Therefore, the frequency of lines in which part of the BAC was inadvertently deleted was 17% (3/18). Expression in two lines with internal deletions, *Bmp4* 5' BAC L8b and *Bmp4* 3' BAC L44, was completely undetectable by XGal stain (data not shown). In addition, pedigree analysis of founder "L8" revealed two independently segregating integrations, with one integration containing all BAC markers (line L8a) and the other integration being fragmented (line L8b). Copy number estimates for the "fragmented" lines indicated each of these lines has an estimated one *lacZ* copy (FIGURE 4.8). In contrast, copy number estimates for breeding lines with "intact" BAC transgenes ranged from >3 to >48 (FIGURE 4.8). In addition, one founder carrying all BAC markers had a copy number estimate of two.

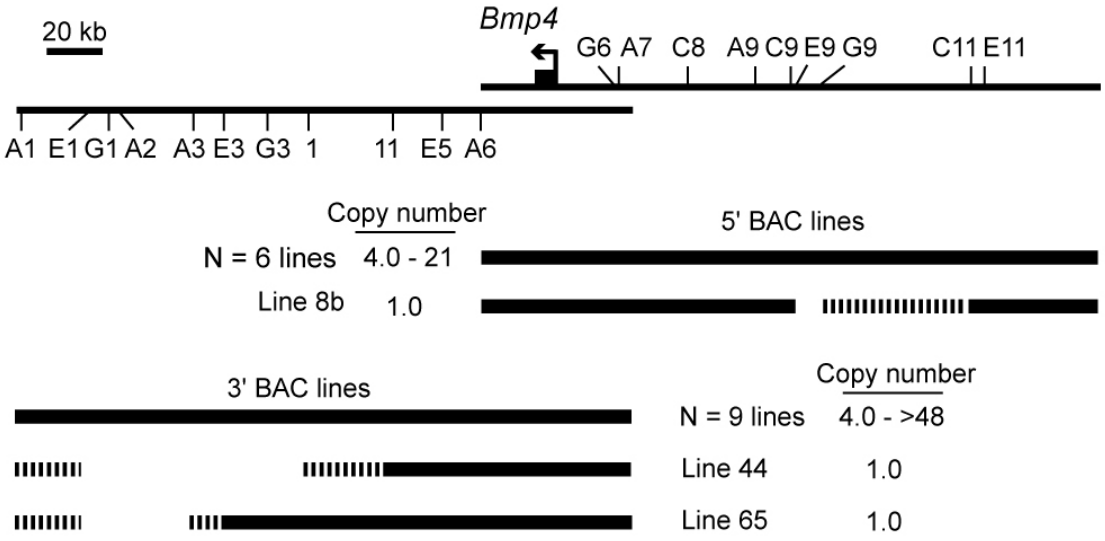

Figure 4.8. Polymorphic marker analysis suggests that transgenic lines that have multiple BAC copies are likely to carry some intact BAC molecules. Polymorphic markers denoted along the length of *Bmp4* 5'BAC (top right) and *Bmp4* 3'BAC (top left) (scale bar=20kb). Shown below each BAC are schematics representing lines for which all BAC markers are present, suggesting intact BAC transgenes (solid bars), and lines containing fragmented BAC transgenes (interrupted bars). Solid lines indicate presence of contiguous transgene-specific markers. Open regions indicate loss of transgene-specific markers, and hatched regions indicate regions of potential breakpoints.

In summary, for the 18 BAC lines or founders that we have analyzed for transgene structure, all those with two or more copies (N=15) appeared to contain intact transgenes based on sampling for STR polymorphisms occurring at an average of every 19 Kb, while those estimated to be single-copy integrants were each missing part of the transgene (N=3). To further investigate the integrity of both low copy and high copy BAC transgenic lines, we performed Southern blot analysis on high molecular weight DNA samples isolated from two individual mice each from a "low copy" *Bmp4* BAC line (5' BAC-L12) and from a "high copy" *Bmp4* BAC line (5' BAC-L1A). Each mouse DNA sample was digested with a rare cutting enzyme (MluI) and subjected to pulsed field gel electorphoresis alongside a digest of purified BAC DNA (FIGURE 4.9a). MluI cuts at two distinct locations in the 5' BAC and digestion of purified 5' BAC DNA yields two bands (FIGURE 4.9a); however, one of the MluI sites is within a CpG island in the *Bmp4* promoter. Therefore, in the context of mouse genomic DNA, it is likely that only the promoter MluI site remains unmethylated and is sensitive to MluI digestion. MluI digestion of transgenic mouse DNA should then yield ~235 kb fragments from 5' BACs integrated as tandem concatamers. After hybridization of the Southern blot with a transgene-specific probe, bands corresponding with the size of a full-length *Bmp4* BAC were visualized (FIGURE 4.9b) suggesting both the low and high copy BAC lines shown here are most likely contain at least one intact molecule. Of note, two *Bmp2* BAC transgene lines previously analyzed by us each had copy numbers of 16 or more and were both shown to contain mostly concatamerized full-length BAC copies by

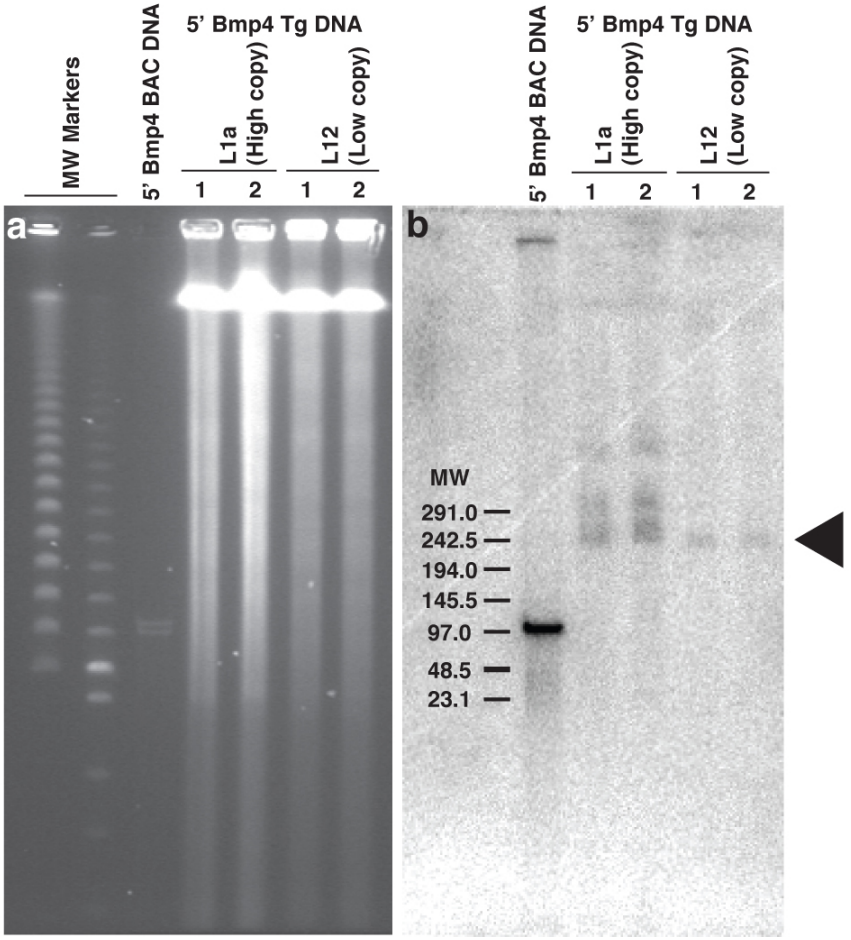

Figure 4.9. Southern blot analysis on high molecular weight DNA samples from low copy (5' L12, avg. copy number = 2) and high copy (5' L1a, avg. copy number = 11) *Bmp4* BAC lines suggests intact transgene copies. (a) Image of ethidium bromide stained pulsed-field gel following electrophoresis of agarose-embedded and MluI-digested high molecular weight DNAs, isolated from embryos generated from 5' BAC carrying stables lines (see Methods). Also included are control digestions of purified 5' BAC DNAs (50 ng DNA per lane). (b) Phosphor-image of gel shown in *A* following Southern transfer and hybridization with radiolabeled probe (IRES-β-geo cassette). The ~110 kb doublet (asterisk) in blot lane 1 represents the expected MluI fragments from purified (unmethylated) 5' BAC DNA. In the lanes with transgenic mouse DNA digested with MluI, bands are evident (arrowhead) that are approximately the full-length size of the 5' BAC transgenes (~235 kb)(note, the high copy line yields stronger bands than the low copy line). This strongly suggests that one or more copies of transgenes are intact in both the high and low copy 5' BAC lines.

Southern blot analysis (Chandler et al. 2007b). Taken together, our data supports the idea that multiple-copy BAC transgene insertions most likely contain one or more full-length BAC copies.

Discussion

BAC transgenic mice are increasingly used for biomedical research. However, few studies have addressed issues regarding BAC transgene copy number, integrity or function across a large population of transgenic mice (Gong et al. 2003) (Alexander et al. 2004). We have taken advantage of the BAC transgenic embryo founders and breeding lines produced in our lab to interrogate the copy number distribution amongst distinct and independent lines, the relationship between copy number and levels of transgene expression, the integrity of multiple BAC transgenes, and the fidelity of BAC transgenes across successive generations.

Several previous reports are also supportive of our observations. The potential for fragmentation of BAC transgenes prior to integration has been noted previously, and can even be used to refine potential *cis*-regulatory domains in some situations (Deal et al. 2006). In a very large set of independently generated BAC transgenic mice (the GENSAT project), Gong *et al.* (Gong et al. 2003) reported that BAC transgene insertions having multiple copies invariably contained full-length copies as tandem arrays, although copy numbers were not reported for individual lines. In some cases for conventional transgenes, stability of transgene copy number has been followed over time in breeding colonies to

document that sporadic loss of transgene copy number can occasionally occur, probably by internal recombinations (Alexander et al. 2004). We have no clear examples of loss of copies within our BAC transgenic lines, although we cannot rule out that it might occur sporadically. Monitoring BAC copy number in each breeding generation seems prudent.

Here, we used a method to obtain copy number information on BAC transgenic lines, using real-time PCR on phenol/chloroform-extracted yolk sac or tail biopsy DNA samples and a BAC copy number standard curve. Unlike the method described here, some previous studies describe real-time PCR methods to estimate copy number on liver biopsy DNA samples, requiring the sacrifice of a transgenic mouse (Ballester et al. 2004). Therefore, our method is advantageous for use with valuable transgenic mice (such as founders) for which premature sacrificing is undesirable. In addition, our method estimates copy numbers based on a standard curve of known BAC copy number standards. Many other studies have used the $2^{-\Delta\Delta Ct}$ method to calculate copy number (Tesson et al. 2002) (Ballester et al. 2004) which requires nearly equivalent PCR efficiencies between the unknown samples and the control. Others have reported that phenol/chloroform-extracted tail DNA samples contain PCR inhibitors (Burkhart et al. 2002) or cannot be quantified accurately by UV spectrophotometry due to phenol contamination (Alexander et al. 2004). However, our studies showed generally reproducible copy number estimates from littermates using tail biopsies prepared in this manner. In some instances, we found samples that gave estimates likely to be erroneous. In these cases,

the estimates were more than two-fold greater or lower than the average for multiple transgenic littermates. This was observed for both tail and yolk sac DNA samples, and we suspect copy number estimates that fell outside of the two-fold threshold were due to phenolic contamination, PCR inhibitors not removed during extraction, or both. Alternatively, loss of transgene copies could result from recombination events in meiosis, as discussed above. This underscores the need to gather, when possible, multiple data points across transgenic littermates and generations for assessment of copy number in individual transgenic lines.

We showed a strong correlation between increased copy number and increased transgene expression in multiple independent transgenic mice derived from three distinct BAC transgene constructs. As copy number increases, so did the qualitative intensity of *lacZ* reporter expression, at least for *Bmp2* and *Bmp4* BACs. Alternatively, others have observed that even for BAC or YAC-sized constructs, increased transgene copy number may not correlate with increased transgene expression and in fact may result in transgene silencing (Heaney and Bronson 2006). Distinct from epigenetic silencing, the sporadic deletion of integrated transgene copies following breeding can clearly reduce transgene expression as compared to expression in preceding generations prior to deletion (Alexander et al. 2004). Although we did not measure transgenic mRNA or beta-galactosidase activity quantitatively, we found no obvious evidence for silencing of *Bmp2* or *Bmp4* BAC transgene expression when copy numbers were high (FIGURE 4.6). We hypothesize lines with higher copy integrations are more

resistant to position effects in comparison to low copy integrations (FIGURE 4.10). We found BAC transgenic lines with at least approximately 10 copies were ideal for analysis, because transgenes were intact and expression of the reporter gene was robust.

Since BAC transgenes are significantly larger constructs compared to conventional transgenes and are more susceptible to fragmentation (Deal et al. 2006), it is imperative that the integrity of BAC transgenes are verified. Here, we show that copy number estimates obtained from real-time PCR methods, coupled with marker genotyping and/or Southern blot analysis, revealed that transgenic lines having at least three copies most likely contain intact molecules as suggested by polymorphic marker analysis, whereas lines with fewer than two copies often contained only partial BAC fragments. We found the majority of our BAC transgene lines likely contained at least one full-length molecule, as demonstrated by polymorphic marker genotyping or Southern blot analysis. This is reassuring due to the general concern that BAC transgenes are easily fragmented. However, we still caution that careful preparation and handling of BAC DNA samples used for pronuclear injection is critical for efficient transgenesis.

Like smaller transgenes, BACs have been suggested to typically incorporate in the genome as 1-5 copy concatamers within a single locus of the genome (Jaenisch 1988) (Giraldo and Montoliu 2001) (Heaney and Bronson 2006). Our data showed that 50% of transgenic lines had between 1-5 copies, and most lines had only single sites of transgene integration. However, we also

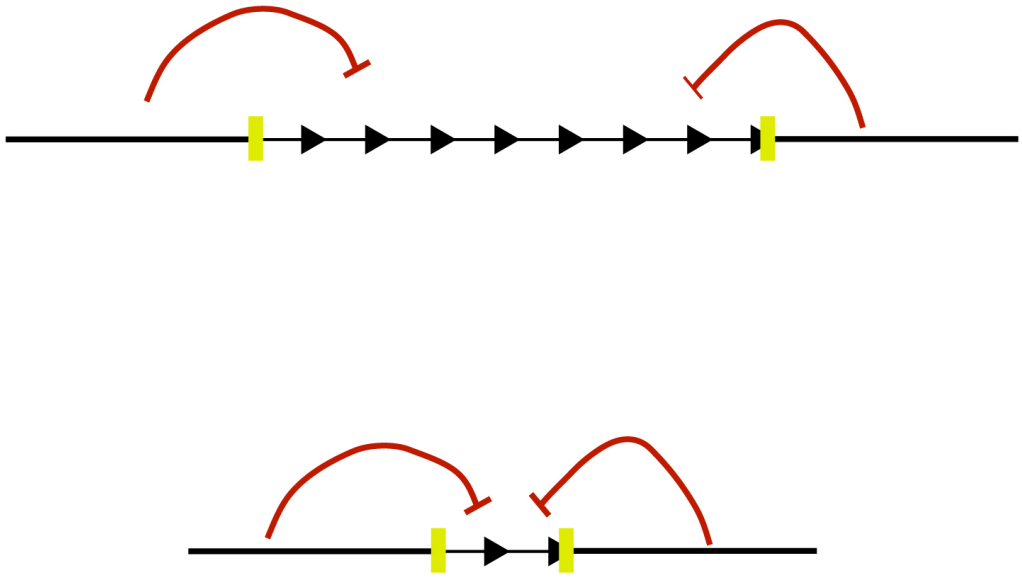

Figure 4.10. Proposed model of high copy versus low copy transgene arrays demonstrates how position effects such as silencing may affect *lacZ* reporter expression. In lines with high copy number estimations, BAC transgenes are likely to be integrated as tandem arrays (top panel). In this instance, any silencing effects may be overcome by the sheer size of the multicopy array with the inner copies protected from position effects. In low copy models, it is likely that a single copy or two integrate and are vulnerable to silencing position effects.

identified multiple lines that each had copy number estimates greater than ~48 or more. Although we are cautious in considering the raw values for these lines as definitive since they were outside the boundaries of our real-time PCR standard curve samples, one line repeatedly produced copy number estimates of nearly 100 (Avg. copy number =96; Std.Dev. =28; n=38) and an additional line with an average copy number of 76 (Std. Dev. =29; n=22). Analysis of a limited number of DNA samples from the latter line (3' BAC-L45A) by dot blot hybridization corroborates this estimate (Avg. copy number =77; n=2). To our knowledge, these are the largest BAC copy number estimates ever reported in transgenic mice.

In addition, three transgenic founders transmitted two independent integration events, suggesting BAC transgenes sometimes integrate in more than one loci of the genome at a reasonably high frequency. Analysis of lines having more than one integration event revealed that the independently segregating integrations often had distinct copy numbers in successive generations. Since we found that increased copy number strongly correlated with increased expression but also that independent integrations often harbor distinct copy numbers, it is imperative that transgenic lines bred for successive generations are carefully characterized by analyzing copy number in multiple F1 progeny to prevent the loss of a valuable integration during subsequent breeding or expression data in F2 progeny. In addition, the analysis of copy number in founders should be approached with caution without confirmation in F1 progeny,

due to the possibility of mosaicism or multiple integration events that cannot be resolved unless the transgene is passed through the germline.

In closing, we present a quick, reliable method for estimating copy number in BAC transgenic lines that has shown to be useful in characterizing multiple lines by analysis of limiting amounts of DNA. These methods are applicable to transient transgenic founder embryos as well where limited tissue is available for DNA extraction (e.g. yolk sacs), albeit with the above caveats. To help reduce costs of reagents, we reduced Taqman reaction volumes from 20 to 10 microliters and used phenol/chloroform extractions to isolate DNA in place of kit-based methods. Finally, we have provided evidence for the importance of evaluating copy number across multiple progeny from BAC transgenic lines, and have demonstrated the increased likelihood that multiple copy integrations typically contain one or more full-length BAC transgenes. We suspect these observations and techniques may be valuable to investigators as the demand for BAC transgenic mice increases.

CHAPTER V

DELETION BAC TRANSGENES SUGGEST ECR2 IS REQUIRED FOR *BMP4* EXPRESSION IN MESODERM

Introduction

Bmp4 is a developmentally regulated gene located on mouse chromosome 14 in a region of the genome that is devoid of other known protein coding sequences. Although the locus is devoid of other genes, a significant amount of conserved noncoding sequence is present throughout this gene desert. Increasing evidence suggests that many conserved noncoding sequences are in fact functional elements of the genome (Wunderle et al. 1998) (Lettice et al. 2002) (Lettice et al. 2003) (Mortlock et al. 2003) (Pennacchio et al. 2006) (Chandler et al. 2007). Ancient, conserved noncoding sequences that are present in fish and mouse genomes often function as transcriptional enhancers (Nobrega et al. 2003) (Kimura-Yoshida et al. 2004) (Woolfe et al. 2005). *In vivo* reporter assays are often used to test DNA for enhancer activity. These experiments can indicate the sufficiency of a DNA segment to direct reporter expression. However, they do not test whether the DNA segment is required for transgene-directed expression. Although deleting a putative enhancer from the endogenous genome is the most definitive way to demonstrate its requirement, this experiment is very expensive and lengthy. An alternative way to test the requirement of an ECR for expression is by deleting the ECR from a reporter transgene.

Comparative analyses have revealed three ancient ECRs present in the *Bmp4* BAC interval of interest (CHAPTERS III and IV). Two ECRs are located in the 5' BAC and one ECR is located in the 3' BAC. Analysis of the 5' and 3' *Bmp4* BAC reporter transgenes *in vivo* revealed numerous sites of expression that were unique to one of the reporter BACs. Therefore, we hypothesized that each ECR may function as an enhancer to direct reporter expression during embryonic development in one of the sites that was unique to the BAC containing the ECR. To test this hypothesis, we engineered three deletion BACs and tested each BAC *in vivo*. Our results indicate ECR 2 is required to direct expression in mesoderm.

Material and Methods

Deletion BAC Reporter Transgenes

Deletion BACs were modified using *galK* selection methods (Warming et al. 2005). SW102 cells were generously provided by Soren Warming and Neil Copeland, NCI, NIH. The 5' and 3' *Bmp4* GFP*lacZ*-BACs (CHAPTER III) were modified to generate three deletion *Bmp4* GFP*lacZ*-BACs. ECR 1 and 2 were deleted from the 5' *Bmp4* GFP*lacZ*-BAC and ECR 3 was deleted from the 3' *Bmp4* GFP*lacZ*-BAC using *galK* homologous recombination methods (Warming et al. 2005).

In brief, SW102 cells were transformed with the 5' or 3' *Bmp4* GFP*lacZ*-BAC and clones were verified by BamHI restriction digestion followed by fingerprint gel analysis. Recombination competent cell preparations were made of SW102 cells containing the 5' or 3' *Bmp4* GFP*lacZ*-BAC. Homology arms

were designed to target each ECR as well as anneal to the *galK* cassette (Table 5.1). Replacement oligonucleotides were designed to create seamless deletions of each ECR by replacing the *galK* cassette with sequence flanking each ECR. PAGE purified homology arms were used to amplify the *galK* targeting cassette. Following PCR amplification with a high-fidelity Taq blend (Expand High Fidelity-Roche), the *galK* targeting cassette was incubated with DpnI restriction enzyme to eliminate plasmid template and gel purified prior to transformation of SW102 recombination competent cells containing the 5' or 3' *Bmp4* GFP*lacZ*-BAC. After recovering cells at 32°C for 1.5 hours, they were washed twice in 1xM9 salts to remove LB from the cells. Transformed SW102 cells were plated on M63 minimal media galactose plates and incubated at 32°C for 3-4 days. Single colonies were isolated and streaked onto MacConkey agar plates with 1% galactose to select for *galK* positive clones. *GalK* positive clones were isolated and verified by restriction digest using a rare-cutting enzyme (MluI) followed by pulsed field gel electrophoresis as previously described (Chapter 3). In addition, fingerprint gel analysis was performed on *galK* positive clones as previously described (Chapter III).

The following *galK* positive clones were selected to make recombination competent cell preparations: Clone 1-2-2 (ECR1 deletion), Clone 2-3-2 (ECR2 deletion), Clone 3-1-3 (ECR3 deletion). PAGE purified replacement oligonucleotides were annealed and used to transform recombination competent *galK* positive clones. Transformed bacterial cells were plated onto M63 minimal

media plates containing 0.2% 2-deoxy-galactose (2-DOG) to select against clones containing the *galK* cassette. Plates were incubated at 32°C for 3 days

Table 5.1. Oligos used for *Bmp4* deletion BAC modifications. Underlined sequence is homologous to the *galK* cassette.

galK deletion oligos	
DelmECR1-F	GGTTTGCCCATTTGGCCAAAGTCACATTCCTTTCGGTGCAAATGCCACCTGTTGACAATTAATCATCGGCA
DelmECR1-R	GGCTTGGTTTCCCTTGCAAGGCTCTTGCCAGCACCTGTGAGCCCTCACCCTCAGCACTGTCCTGCTCCTT
DelmECR2-F	CAGCCCTGAGTAACAGAGAGAGGGAAGGCAGGAGGTTAAACCAAACTGTTCCTGTTGACAATTAATCATCGGCA
DelmECR2-R	GAGAAGCTCTGCTTCCCAAAGTTCCCTACATAATCCTTACCGTGAAGAGCTCAGCACTGTCCTGCTCCTT
DelmECR3-F	TAAAGCAAAGACCTGTGCTGTGAGCCAGAGCTGATCACAAGATCAAAGCCCCTGTTGACAATTAATCATCGGCA
DelmECR3-R	ACATTATTCAACAAACAAAACACTCTCATTCTAAAAGAGAAAGAAAAAAATCAGCACTGTCCTGCTCCTT
galK replacement oligos	
RepmECR1-F	GGTTTGCCCATTTGGCCAAAGTCACATTCCTTTCGGTGCAAATGCTGCCAGGGTGAGGGCTCACAGGTGCTGGCAAGAGCCTTGCAAGGGAAACCAAGCC
RepmECR2-R	GGCTTGGTTTCCCTTGCAAGGCTCTTGCCAGCACCTGTGAGCCCTCACCCTGGCAGCATTTGCACCGAAAGGAATGTGACTTTGGCCAAATGGGCAAACC
RepmECR2-F	CAGCCCTGAGTAACAGAGAGAGGGAAGGCAGGAGGTTAAACCAAACTGTTGCTCTTCACGGTAAGGATTATGTAGGGAACTTTGGGAAGCAGAGCTTCTC
RepmECR2-R	GAGAAGCTCTGCTTCCCAAAGTTCCCTACATAATCCTTACCGTGAAGAGCAACAGTTTGGTTTAACCTCCTGCCTTCCCTCTCTCTGTTACTCAGGGCTG
RepmECR3-F	TAAAGCAAAGACCTGTGCTGTGAGCCAGAGCTGATCACAAGATCAAAGCCTTTTTTTCTTTCTCTTTTAGAATGAGAGTGTTTTGTTTGTTGAATAATGT
RepmECR3-R	ACATTATTCAACAAACAAAACACTCTCATTCTAAAAGAGAAAGAAAAAAAGGCTTTGATCTTGTGATCAGCTCTGGCTCACAGCACAGGTCTTTGCTTTA

and colonies were isolated and grown to prepare minicultures. Crude alkaline lysis preparations were made using minicultures and clones were verified by pulsed field gel electrophoresis and fingerprint gel analysis as described above. NucleoBond® AX 500 (Clontech) kits were used to purify BAC DNA for sequencing. The following primers were designed outside the deleted ECR and used to sequence across the deleted segment of the BAC: Deletion 1, 5'-AGGACTAGGGTTTGCCCATT-3'; Deletion 2, 5'-CTCCAGGCTCAGATGTGGTT-3'; Deletion 3, 5'-GCCAAAATACCCGTGTGACT-3'. The following clones were selected for cesium chloride purification, gel quantitation and pronuclear injection as described previously (see Chapter 3): Clone 1-2-2-4 (Deletion 1), Clone 2-3-2-2 (Deletion 2), Clone 3-1-3-4 (Deletion 3).

Transgene Expression Analysis

Transgenic lines were established from founder mice and *lacZ* expression was analyzed at 9.5, 12.5, and 15.5 dpc for all lines as previously described (Chapter III). Lines were examined for transgene integrity by polymorphic marker analysis and/or copy number estimation as previously described (see Chapter IV). Lines with low or undetectable *lacZ* expression and/or fragmented transgenes were excluded from further analysis.

Embryo Processing and Imaging

Embryos were processed and imaged as previously described (Chapter III). Histological analysis was performed as previously described (Chapter III). Founder mice were identified using PCR-based genotyping as described previously (Chapter III).

Results

Deletion BAC Transgenes

To test the requirement of each ECR for reporter expression driven by the 5' or 3' GFP*lacZ*-BAC, three deletion BACs were engineered and used to generate transgenic mouse lines. Approximately 200 bp deletions were made in the 5' or 3' GFP*lacZ*-BACs (Chapter III) using the *galK* selection method of homologous recombination (Warming et al. 2005). This method provides a robust and efficient way to engineer deletions without leaving any remaining exogenous sequence in place of the deletion. Figure 5.1 depicts the location of each deletion on the UCSC Genome Browser (May 2004 Assembly) as a hatched black bar. Homology arms were designed to remove each ECR as identified by fish/mouse sequence comparisons (Chapter II) as well as approximately 50 bp of additional sequence flanking each ECR (solid red bar, FIGURE 5.1). Deletion 1 and Deletion 2 were engineered in the 5' GFP*lacZ*-BAC and Deletion 3 was engineered in the 3' GFP*lacZ*-BAC (FIGURE 5.1). Deletions ranged from 180-221 bp in size. Note, the amount of conservation between mammalian species (Rat/Human/Dog/Chicken Multiz Alignments & PhyloHMM

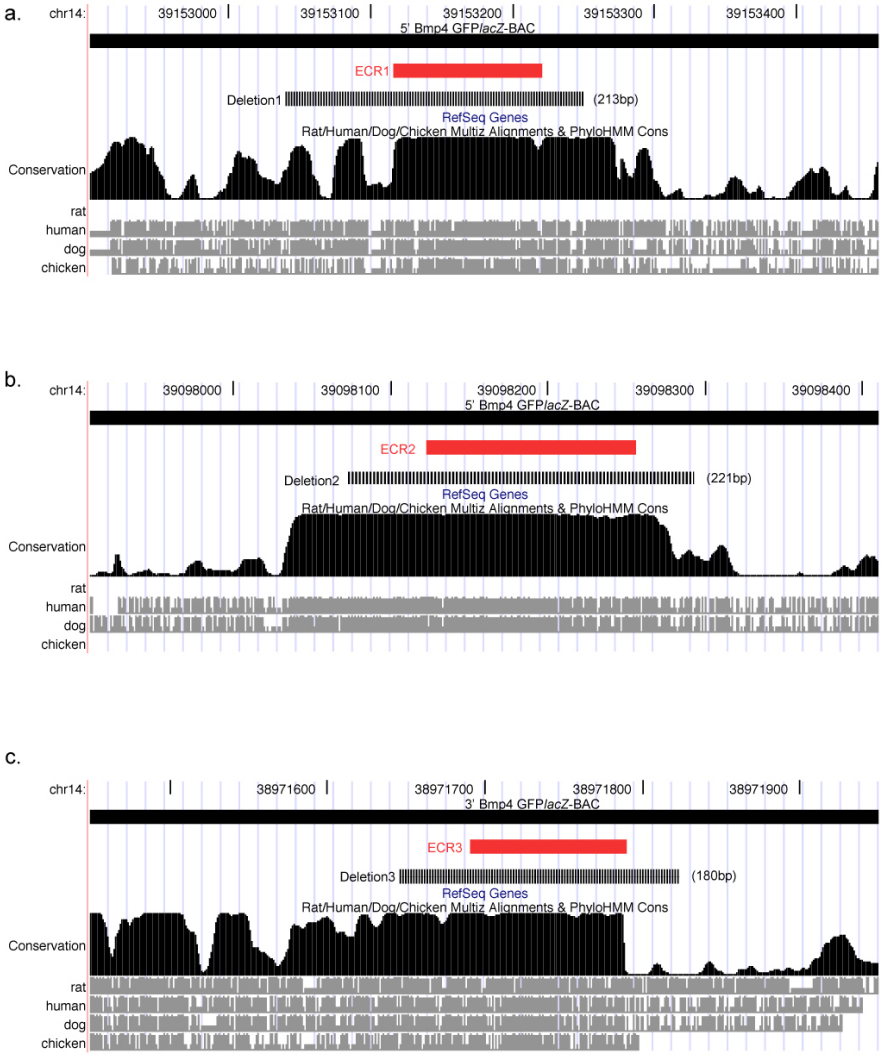

Figure 5.1. *Bmp4* fish/mouse ECRs as well as the region deleted from the 5' or 3' GFP*lacZ*-BACs is depicted in these UCSC Genome Browser (May 2004 Assembly) plots. (a) ECR1 and Deletion 1 are depicted below the 5' GFP*lacZ*-BAC. (b) ECR2 and Deletion 2 are depicted below the 5' GFP*lacZ*-BAC. (c) ECR3 and Deletion 3 are depicted below the 3' GFP*lacZ*-BAC. Note, each deletion (hatched, black bar) is approximately 200 bp in size and completely eliminates the entire fish/mouse ECR (red bar). The Conservation track (filled graph) depicts the amount of sequence conservation in cross species alignments (Siepel et al. 2005).

Cons Track) extends beyond the ECRs identified by fish/mouse sequence comparisons (FIGURE 5.1). However, each Deletion encompasses the entire ECR identified by fish/mouse sequence comparisons (FIGURE 5.1). While comparisons to chick or mammals may indicate broader regions of less ancient conservation, we were specifically interested in testing the requirement of the core regions of mammal/fish homology. Prior to pronuclear injection, each purified Deletion BAC was verified to contain the desired modification and no rearrangements by restriction enzyme digestion followed by pulsed field gel electrophoresis and fingerprint gel analysis as described in Chapter III. Large rearrangements or deletions aberrantly made could be resolved in this manner. In addition, the degree of DNA degradation was visible upon gel electrophoresis. BAC DNA prepared using a cesium chloride gradient resulted in higher quality DNA as seen by the lack of smearing in the high molecular weight range (FIGURES 5.2 and 5.3). No aberrant rearrangements or deletions were detected as shown by the banding pattern of the Deletion BACs that mirrored the 5' or 3' GFP*lacZ*-BAC banding pattern (FIGURES 5.2 and 5.3). Since gel electrophoresis of BAC DNA could not detect small deletions, each Deletion BAC was sequenced across the deleted segment to verify each ECR had been removed (data not shown). Deletion BACs were purified over a cesium chloride gradient and used for pronuclear injections as intact, circular molecules to generate transgenic mouse lines.

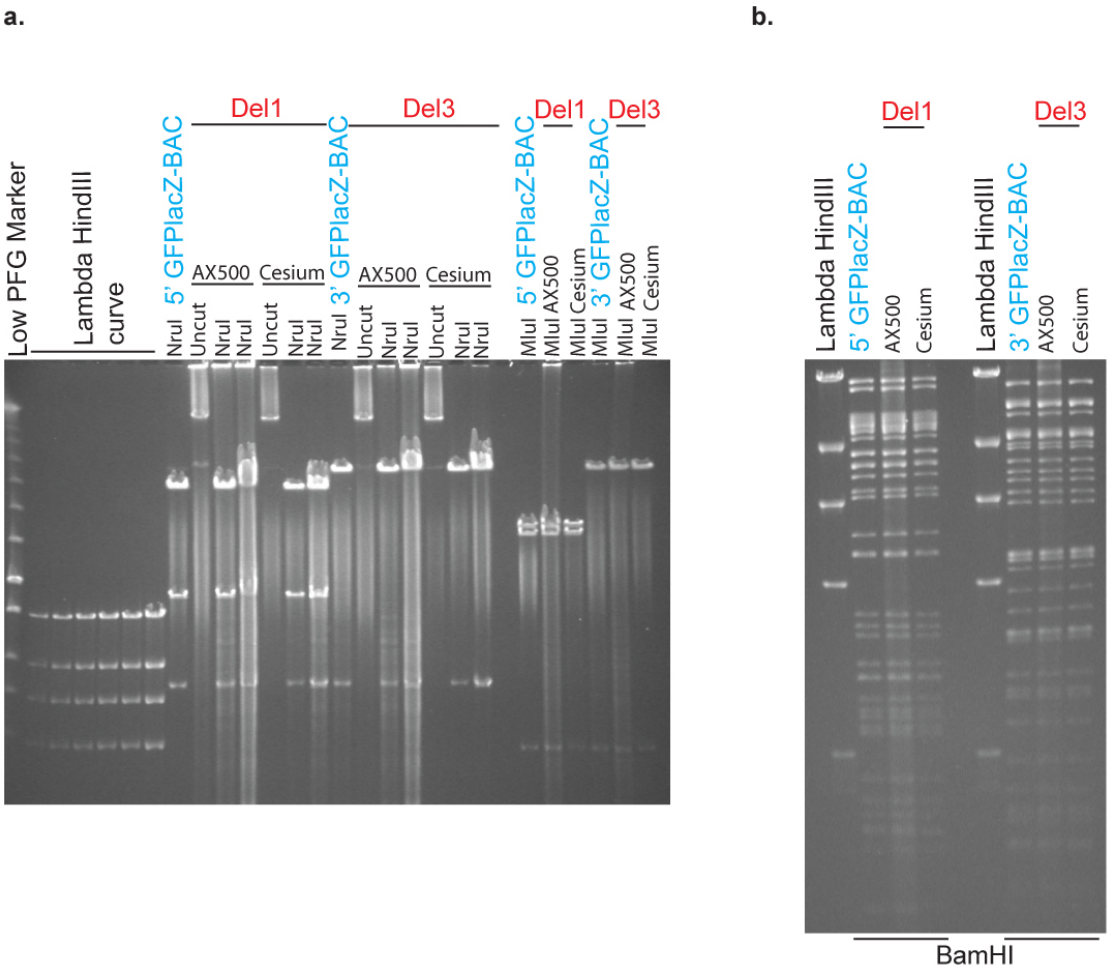

Figure 5.2. Analysis of Deletion BAC quality and structure reveals Deletion 1 and Deletion 3 BACs are without aberrant deletions or rearrangements. (a) Pulsed field gel electrophoresis of Deletion 1 and Deletion 3 BAC DNA. (b) Fingerprint gel electrophoresis of Deletion 1 and Deletion 3 BAC DNA. Restriction digests of BAC DNA prepared using an AX500 kit (Clontech) are run alongside BAC DNA prepared over a cesium chloride gradient. Restriction digests of 5' and 3' GFP*lacZ*-BAC DNA serve as controls. NruI and MluI rare-cutting restriction enzymes were used to digest an AX500 and cesium chloride gradient prepared Deletion BACs followed by pulsed field gel electrophoresis. The frequently-cutting restriction enzyme, BamHI, was used to digest BAC DNA preparations which were then subject to fingerprint gel analysis as described in Chapter III.

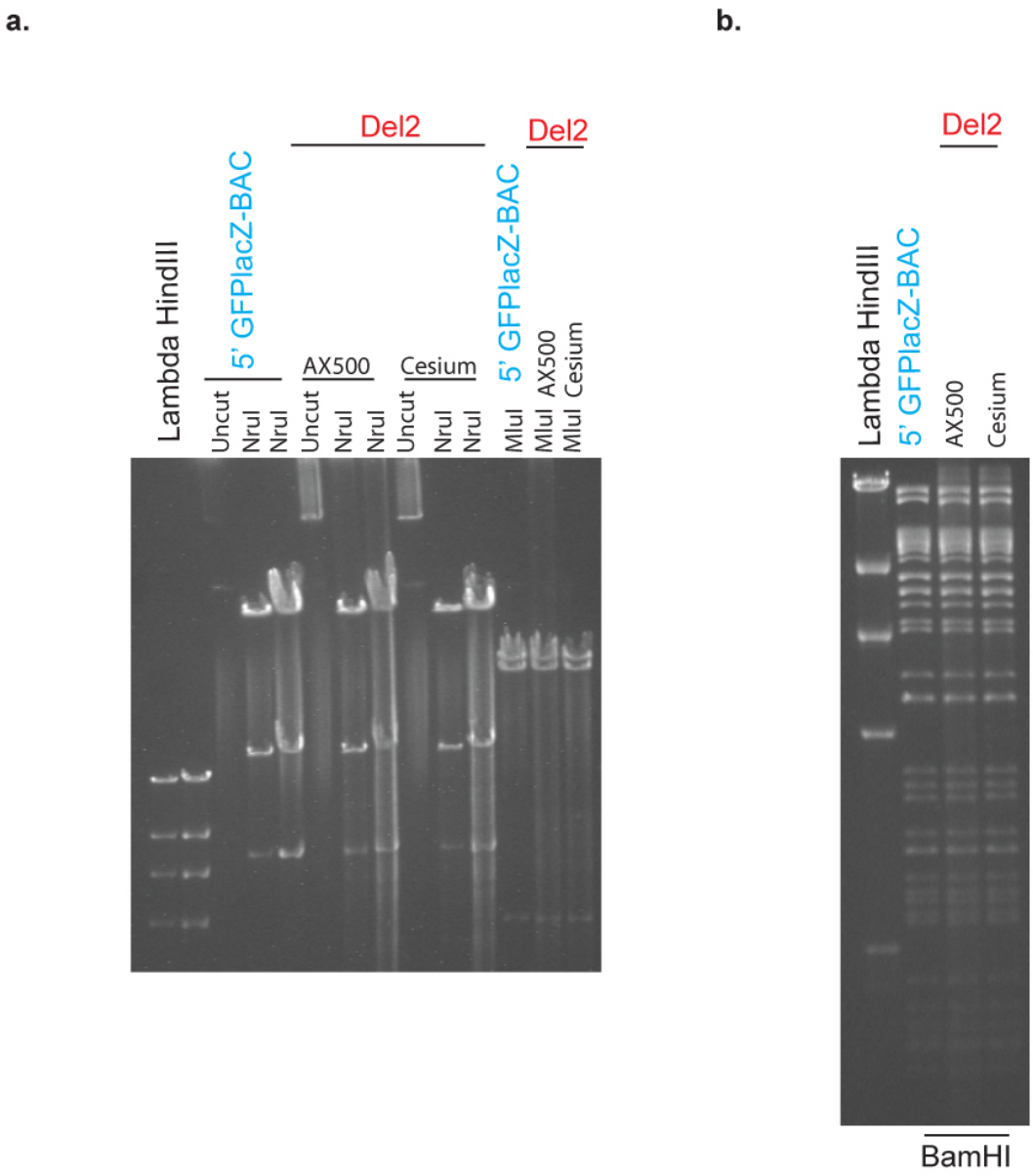

Figure 5.3. Analysis of Deletion BAC quality and structure reveals Deletion 2 BAC is without aberrant deletions or rearrangements. (a) Pulsed field gel electrophoresis of Deletion 2 BAC DNA. (b) Fingerprint gel electrophoresis of Deletion 2 BAC DNA. Restriction digests of BAC DNA prepared using an AX500 kit (Clontech) are run alongside BAC DNA prepared over a cesium chloride gradient. Restriction digests of 5' GFP*lacZ*-BAC DNA serve as controls. NruI and MluI rare-cutting restriction enzymes were used to digest an AX500 and cesium chloride gradient prepared BAC DNA followed by pulsed field gel electrophoresis. The frequently-cutting restriction enzyme, BamHI, was used to digest BAC DNA preparations which were then subject to fingerprint gel analysis as described in Chapter III.

Single Founder Generated with Deletion 1 BAC Suggests ECR1 is Not Required for *Bmp4* Expression During Development

Once Deletion BAC lines were established, embryos were generated and stained with Xgal to detect *lacZ* expression. To obtain a snapshot of expression throughout development, embryos were generated at 9.5, 12.5 and 15.5 dpc. These timepoints were chosen because they represented the onset, middle and end of organogenesis. Staining patterns in deletion BAC embryos were then compared to full-length 5' or 3' GFP*lacZ*-BAC embryos (see Chapter III) to assess any potential changes in reporter expression with the removal of each ECR. A single founder was obtained from pronuclear injections of Deletion BAC 1 and a breeding line was established. The Deletion 1 BAC line (L20) exhibited robust reporter expression (data not shown) and high copy number (Neo avg = 9) (Chapter IV). Therefore, given our results in Chapter IV, this line most likely contains an intact transgene. When compared to the full length 5' BAC, *lacZ* expression patterns were similar and no obvious differences, such as ectopic or missing expression patterns, were noticed. Unfortunately, only one founder was identified for the Deletion 1 BAC transgene. Since it is be necessary to repeat a transgenic result in at least two independent transgenic lines, Deletion 1 BAC should be reinjected to make conclusive statements. However, we hypothesize ECR1 is not required for *Bmp4* expression. Nevertheless, data from this line corroborates the expression patterns seen in embryos generated from the full-length 5' BAC line, lending weight to our findings in Chapter III and arguing that ECR1 is dispensable for embryonic *Bmp4* expression.

Deletion 3 BAC Fails to Elucidate a Role for ECR3 in the Expression of *Bmp4*

Four founders were identified from pronuclear injections of Deletion 3 BAC allowing lines to be propogated (L15, L19, L20, L22). Embryos were generated at 12.5 and 15.5 dpc for Xgal staining to detect *lacZ* expression. Since the full-length 3' BAC did not direct reporter expression at 9.5 dpc (FIGURE 3.8), embryos were not obtained at that stage. *lacZ* expression was not detected in embryos generated from 2/4 lines (L19, L20). Likewise, copy number estimates were low for both lines (L19, Neo= 1; L20, Neo= 2) (see CHAPTER IV). However polymorphic marker analysis suggested both lines contained intact transgenes (see CHAPTER IV), suggesting low or undetectable expression levels may be due to silencing position effects or an undetected transgene fragmentation (see Chapter IV). Therefore, L19 and L20 were excluded from further analysis since loss of expression could not definitively be attributed to the intentional deletion. Polymorphic marker analysis and copy number estimation suggested L15 and L22 contained intact BAC transgenes (Chapter IV). In addition, copy number estimates for each line were high (L15, Neo=21; L22, Neo=10) (CHAPTER IV). Thus, embryos generated from these lines were compared to age-matched embryos generated from the 3' GFP*lacZ*-BAC (FIGURE 5.6). Embryos generated from Deletion 3 BAC lines at 12.5 and 15.5 dpc showed no obvious differences from the control (3' GFP*lacZ*-BAC) (FIGURE 5.6). For example, 12.5 dpc Deletion 3 BAC embryos displayed expression in dorsal root ganglia, proximal limb, and craniofacial mesenchyme much like age-matched embryos generated from the 3' GFP*lacZ*-BAC (FIGURE 5.6). This trend

continued at 15.5 dpc when Deletion 3 BAC embryos exhibited the same sites of expression as 3' GFP*lacZ*-BAC embryos (FIGURE 5.6). These results suggest that the loss of ECR3 does not result in the loss of expression patterns seen at 12.5 and 15.5 dpc.

Deletion 2 BAC Reveals a Critical Role for ECR2 in Expression of *Bmp4* in Posterior Lateral Plate Mesoderm

Four founders were identified for Deletion 2 BAC and five lines were propagated (L7, L8a, L8b, L9, L13), since one founder gave rise to two lines (CHAPTER IV). Lines with intact transgenes as suggested by polymorphic marker analysis and copy number estimation (CHAPTER IV), as well as robust lacZ expression were identified for further analysis. Polymorphic marker analysis and copy number estimation suggested 4/5 (L7, L8a, L9, L13) lines were each most likely intact. However, polymorphic marker analysis revealed L8b was fragmented (Chapter IV) and Xgal staining resulted in no *lacZ* expression in multiple transgenic embryos, which prevented the use of this line for deletion analysis. Although polymorphic marker analysis suggested L9 and L13 contained intact transgenes, Xgal staining of transgenic embryos revealed very low levels of expression (FIGURE 5.4). In addition, copy number estimates were low (L9, Neo=5; L13, Neo=3). Therefore, these lines were excluded from detailed analysis since absent expression patterns could not be definitively linked to the engineered deletion. Both L7 and L8a had high copy number estimations (L7, Neo= 21; L8a, Neo>48) and robust *lacZ* expression. Polymorphic marker

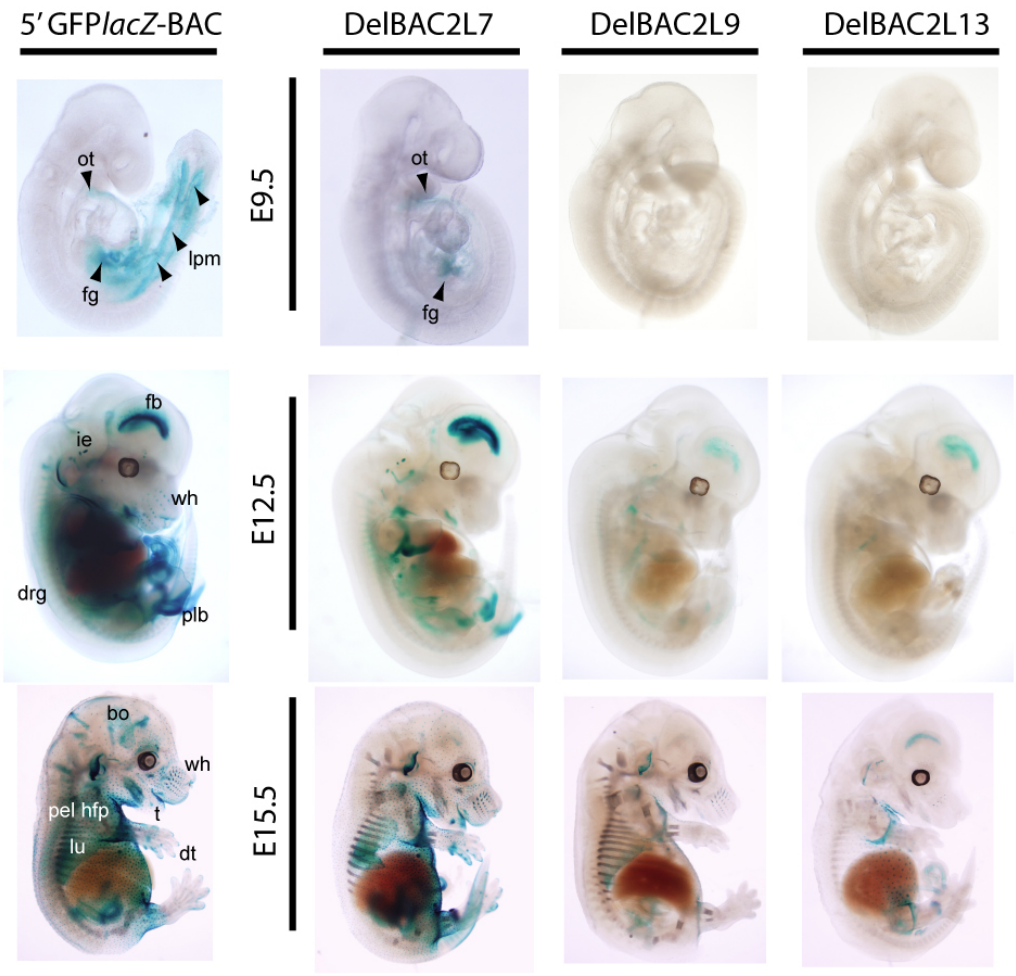

Figure 5.4. Analysis of *lacZ* expression in Deletion 2 BAC embryos compared to 5' GFP*lacZ*-BAC embryos reveal a loss of expression in posterior mesoderm at 9.5 dpc. Xgal stained embryos from 3/4 Deletion 2 BAC lines were generated at 9.5, 12.5 and 15.5 dpc and compared to age-matched 5' GFP*lacZ*-BAC Xgal stained embryos. Note, *lacZ* expression in mesoderm of the 5' GFP*lacZ*-BAC embryo at 9.5 dpc (arrows). Mesoderm expression is absent in Deletion 2 BAC embryos, but present in foregut and heart (L7). *LacZ* expression in later stage Deletion 2 BAC embryos (12.5, 15.5 dpc) is similar to control (5' GFP*lacZ*-BAC). Note, Deletion 2 BAC L9 and L13 have very weak and/or undetectable expression (last two panels). Therefore, these lines were not used to analyze the requirement of ECR2 for *lacZ* expression, although some structures had faint staining in patterns that matched the full-length 5' BAC. (ot= outflow tract, fg= foregut, lpm= lateral plate mesoderm, fb= forebrain, ie= inner ear, drg= dorsal root ganglia, wh= whiskers, plb= posterior limb bud, bo= bone, pel hfp= pelage hair follicle placodes, lu= lung, t= tooth, dt= digit tips)

analysis suggested both Deletion 2 BAC lines contained intact transgenes, narrowing the chance that loss of expression was due to an aberrant deletion of part of the transgene. Therefore, these lines were used for detailed analysis.

Deletion 2 L7 9.5 dpc embryos revealed dramatic loss of expression in posterior lateral plate mesoderm (FIGURE 5.4), whereas age-matched 5' GFP*lacZ*-BAC embryos clearly had reporter expression in posterior lateral plate mesoderm (FIGURE 5.4, arrows). The loss of expression in Deletion 2 BAC embryos could not be explained by low copy number estimates since they were higher than copy number estimations for the full-length BAC (FIGURE 5.5). Other sites of expression that are directed by the 5' GFP*lacZ*-BAC at 9.5 dpc such as heart and foregut are present in the Deletion 2 embryos (FIGURE 5.4, arrows) indicating the loss of ECR2 specifically results in the loss of expression in posterior lateral plate mesoderm at 9.5 dpc. Reporter expression in 12.5 and 15.5 dpc Deletion 2 embryos failed to differ from control (5' GFP*lacZ*-BAC) embryos (FIGURE 5.4). Embryos generated at 10.5 dpc from Deletion 2 BAC L7 and L8a substantiated the loss of expression in posterior lateral plate mesoderm seen at 9.5 dpc (FIGURE 5.5). In both lines, Deletion 2 embryos displayed robust expression similar to control embryos in structures such as the developing limbs and forebrain at 10.5 dpc (FIGURE 5.5). However, expression in lateral plate mesoderm was abolished in Deletion 2 embryos (FIGURE 5.5, arrows). In addition, although expression in the heart was observed in at least 1 9.5 dpc Deletion 2 embryo (see L7, Figure 5.4), no expression in the heart was observed at 10.5 dpc for this line or line L8a (Figure 5.5, asterisks) Therefore, the loss of

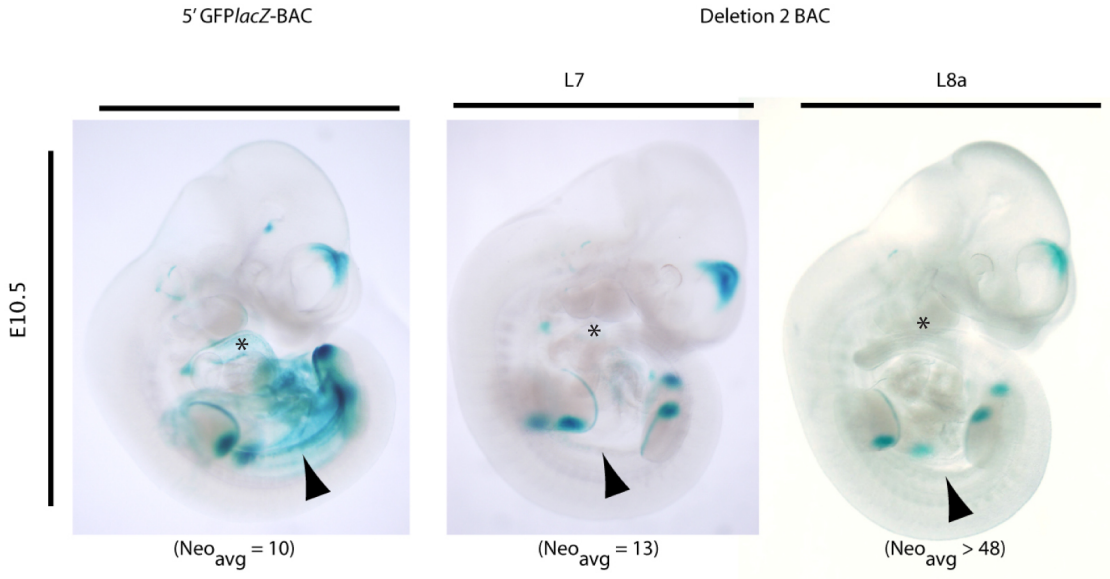

Figure 5.5. Loss of mesoderm expression is reproducible in independent Deletion 2 BAC transgenic lines. Xgal stained embryos from Deletion 2 BAC L7 and L8a at 10.5 dpc show loss of *lacZ* expression in lateral plate mesoderm (arrows), yet most other sites of 5' GFP*lacZ*-BAC-driven expression remain. In both lines loss of *lacZ* expression was also observed in the heart (asterisks) at 10.5 dpc. Copy number estimates are shown as averages for each line.

lacZ expression in Deletion 2 BAC embryos is most likely due to the deletion of ECR 2 suggesting ECR2 is critical for *Bmp4* expression in lateral plate mesoderm at 9.5-10.5 dpc.

The 5' GFP*lacZ*-BAC directed expression in extraembryonic mesoderm at 7.5 dpc (Chapter III). Following the above analysis of the Deletion 2 BAC, further tests of ECR2 sequences in minigene assays suggested it had enhancer function in the extraembryonic mesoderm at this stage (see Chapter VI). Therefore, we examined Deletion 2 embryos at 7.5 dpc. The loss of ECR2 partially ablated expression in the extraembryonic mesoderm, but not completely (data not shown). More specifically, Deletion 2 BAC embryos retained lacZ expression in the allantoic bud portion of the extraembryonic mesoderm, but not in the chorionic or amnionic mesoderm at 7.5 dpc.

Discussion

Analysis of *lacZ* expression in 5' and 3' GFP*lacZ*-BAC embryos strongly suggested multiple long-range enhancers exist to impart spatiotemporal specific expression of *Bmp4* throughout development (Chapter III). In addition, comparative analysis of genomic sequence surrounding *Bmp4* in fish and mouse found three ancient noncoding sequences located far from the *Bmp4* promoter (Chapter II). Therefore, we hypothesized these three sequences functioned as tissue-specific enhancers. To help test this hypothesis, we deleted each ECR from its respective GFP*lacZ*-BAC and tested the Deletion BACs in vivo. One of

these three Deletion BACs clearly demonstrated a functional role for the deleted ECR.

ECR2 is required for normal expression of *Bmp4* in lateral plate mesoderm at 9.5-10.5 dpc. Additionally, expression was reduced in extraembryonic mesoderm at 7.5 dpc. Incomplete ablation of extraembryonic mesoderm expression may be due to the design of the deletion since additional noncoding conservation exists beyond the borders of the fish/mouse ECR that was deleted (FIGURE 5.1). Alternatively, extraembryonic mesoderm is subdivided into distinctly specified tissues (amnionic, chorionic, allantoic mesoderm) (Hogan 1994) that may allow ECR2 to direct expression in the chorionic and amnionic portion of extraembryonic mesoderm, but not the allantoic portion. However, this is would be surprising since *Bmp4* is expressed throughout all three portions of extraembryonic mesoderm (Lawson et al. 1999). It may be useful to delete a much larger segment containing the fish/mouse/ human ECR2 as well as additional flanking sequence to rule out a requirement for the inter-mammal conserved sequences flanking ECR2 in portions of extraembryonic mesoderm.

Alternatively, complete expression in the extraembryonic mesoderm may require multiple *cis*-regulatory elements and ECR2 represents one of the elements. These modular elements may be required for the separate extraembryonic mesoderm domains (chorionic, amnionic, allantoic). Or, each element may be partially redundant with the other element(s) directing some expression throughout the extraembryonic mesoderm, yet full strength of

expression is achieved with the coordinate efforts of each element. In fact, others have suggested redundant *cis*-regulatory elements exist for *Shh* (Jeong et al. 2006). In this study, Jeong and colleagues determined two *cis*-regulatory elements directed expression in the hindbrain and spinal cord (Jeong et al. 2006). When each element was independently deleted from a BAC transgene, expression persisted in hindbrain and spinal cord (Jeong et al. 2006). However, when both elements were deleted from a single BAC transgene, expression was nearly undetectable in hindbrain and spinal cord (Jeong et al. 2006). Taken together, these results suggest *Shh* expression in hindbrain and spinal cord are coordinated by two independent *cis*-regulatory elements. To investigate the possibility that ECR2 is a redundant *cis*-regulatory element, additional deletions across the 5' BAC could be engineered and tested for reporter activity in extraembryonic mesoderm. If an additional region beyond ECR2 is sufficient for extraembryonic mesoderm expression, comparative analysis may be used to identify candidate enhancers and these regions could be specifically tested for extraembryonic mesoderm activity.

Deletion BAC experiments suggest ECR2 is required for both embryonic (lateral plate mesoderm) and normal extraembryonic mesoderm expression of *Bmp4*. Extraembryonic and embryonic mesoderm arise from a common source (inner cell mass, ICM) that exists prior to implantation of the fertilized egg at 3.5 dpc (Hogan et al. 1994) (see FIGURE 1.5). Likewise, *Bmp4* is expressed in the ICM at 3.5 dpc (Coucouvanis and Martin 1999). Thus, we cannot rule out the requirement of ECR2 for directing *Bmp4* expression in ICM. However, this is

unlikely since the ICM also gives rise to endoderm and ectoderm. Therefore, if ECR2 was an ICM enhancer and we observed reporter expression in extraembryonic mesoderm and lateral plate mesoderm as a result of lineage tracing the ICM cells, we would also expect to see reporter expression in endoderm and ectoderm. Future studies of 5' GFP*lacZ*-BAC and Deletion 2 BAC *lacZ* expression in the ICM of preimplantation embryos may address this hypothesis.

Deletion 1 and 3 BAC mice revealed ECR1 and 3 are not required for *Bmp4* BAC transgene expression at 9.5, 12.5 or 15.5 dpc. Although each ECR is highly conserved, it is possible that they are not *Bmp4 cis*-regulatory elements. A more likely explanation for the failure to detect a requirement for ECR3 lies in the design of the Deletion BACs. Each ECR identified by fish/mouse sequence comparisons was deleted from the full length BACs (FIGURE 5.1). However, a considerable amount of conservation amongst mammals persists beyond the confines of each fish/mouse ECR. Therefore, it is possible that critical transcription factor binding motifs sufficient for ECR3 to function as a tissue-specific enhancer are present in the flanking segments conserved amongst mammals. In the future, it would be beneficial to engineer larger deletions in the reporter BACs to test the requirement of ECR 1 and 3. Alternatively, ECR1 and ECR3 may function redundantly with other unknown enhancers (or each other) not tested in our assay. Also, Deletion 1 or 3 BAC embryos may show a requirement for ECR 1 or 3 to direct a site of expression not present at the timepoints that were analyzed (eg. adult). In this regard, it would be

advantageous to generate embryos at 10.5 dpc, since it is an intermediate stage between the other timepoints assayed (9.5, 12.5 dpc) and both full-length BACs exhibit *lacZ* expression at 10.5 dpc (FIGURE 3.8). While this deletion analysis cannot test all potential hypotheses, it serves as a useful tool for identifying critical, non-redundant *cis*-elements.

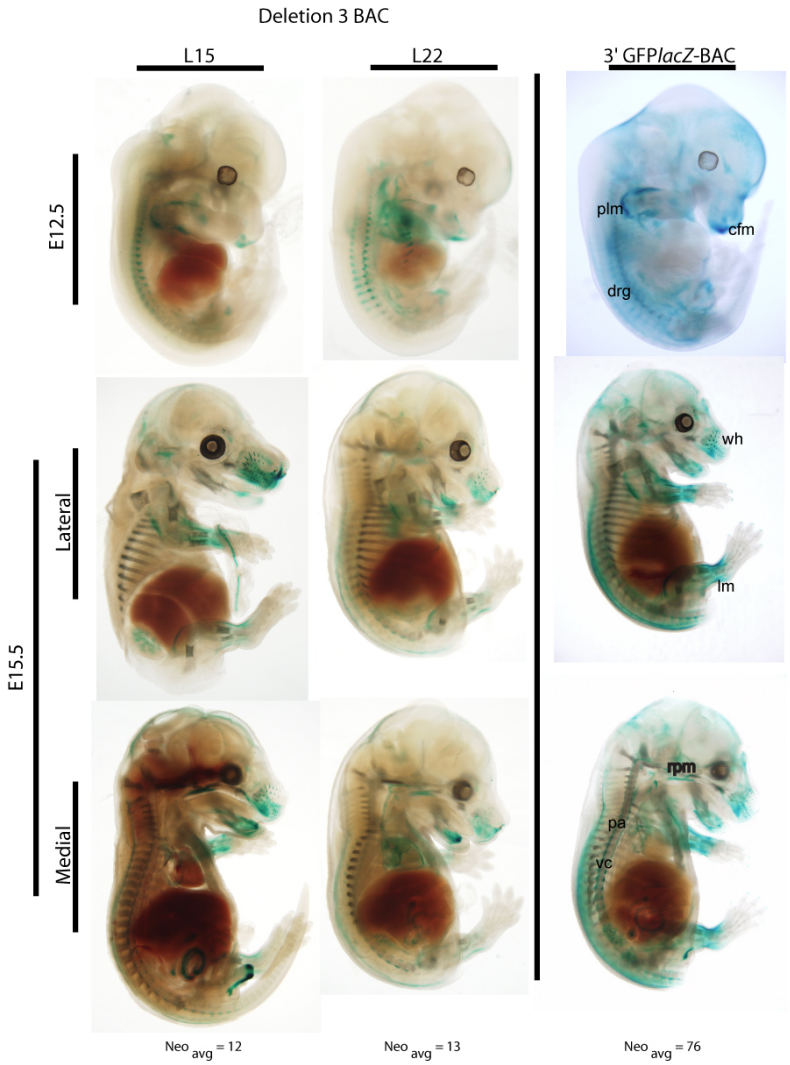

Figure 5.6. Analysis of *lacZ* expression in Deletion 3 BAC embryos compared to 3' GFP*lacZ*-BAC embryos at 12.5 and 15.5 dpc fail to demonstrate a loss of tissue-specific expression. Xgal stained embryos from 2/4 Deletion 3 BAC lines were generated at 12.5 and 15.5 dpc and compared to age-matched 3' GFP*lacZ*-BAC Xgal stained embryos. Expression could not be detected in transgenic embryos generated by L19 and L20 and are not shown. Note, embryos at 12.5 dpc as well as medial and lateral views of 15.5 dpc embryos generated from independent Deletion 3 BAC lines mimic expression seen in age-matched 3' GFP*lacZ*-BAC embryos. Note, *lacZ* expression is present in the along the dorsal portion of each embryo as seen in the 3' BAC, but may appear missing due to bisection of the embryo. Copy number estimates are shown as averages for each line. (plm=proximal limb mesenchyme, cfm= craniofacial mesenchyme, drg= dorsal root ganglia, wh= whisker, lm= limb mesenchyme, rpm= roof palate mesenchyme, pa= pulmonary artery, vc= vertebral column)

CHAPTER VI

ECRS EXHIBIT ENHANCER ACTIVITY IN TRANSGENIC FISH ASSAY

Introduction

Significant efforts over the past two decades have allowed researchers to locate, annotate, and study the function of protein coding sequences in vertebrate and invertebrate genomes alike. Now that the majority of genes have been found, efforts have shifted towards locating and understanding the noncoding functional elements peppered throughout the vast genomic expanse. Studies have shown that many noncoding sequences conserved between teleosts and mammals function as tissue-specific enhancers (Nobrega et al. 2003) (Pennacchio et al. 2006) (Woolfe et al. 2005) (Kimura-Yoshida et al. 2004) (Goode et al. 2003) (Ghanem et al. 2003) (Barton et al. 2001). Comparative analyses revealed three long-range noncoding ECRs flanking *Bmp4* in the mouse and pufferfish genomes (see Chapter II). Although pufferfish is an excellent model organism for comparative genomics, it is not amenable to genetic experimentation. Alternatively, the zebrafish is an attractive model organism for testing enhancer activity because 1) development occurs in the external environment allowing continuous observation of the same embryos without sacrificing, 2) development is rapid, 3) transparent embryo facilitates GFP visualization, 4) transient transgenic analysis is cost effective, and 5) microinjection technique is simple and straightforward. The main disadvantage of transient analysis in transgenic zebrafish embryos is mosaic expression

(Stuart et al. 1990) (Muller et al. 2002). More specifically, transgenes that have not been passed through the germline (transient analysis) are not present in all cells leading to highly mosaic expression in the structure that is labeled.

Zebrafish *Bmp4* was cloned subsequent to the mouse and human genes (Martinez-Barbera et al. 1997). Despite a round of whole-genome duplication after the divergence of teleost fish from other vertebrates (Canestro et al. 2007), only one copy of zebrafish *Bmp4* has been discovered (Martinez-Barbera et al. 1997). The amino acid sequence in the mature region of zebrafish Bmp4 is highly conserved with mouse Bmp4 (92%) (Martinez-Barbera et al. 1997). *Bmp4* transcripts are first detected by RT-PCR at 4 hours post fertilization (hpf) (Martinez-Barbera et al. 1997) and by whole mount *in situ* hybridization at 5 hpf (Nikaido et al. 1997). Northern blot analysis revealed *Bmp4* is not maternally expressed (Nikaido et al. 1997). Upon gastrulation (6hpf), *Bmp4* expression is robust in the ventral domain of the embryo as well as in the embryonic shield (Nikaido et al. 1997). As gastrulation proceeds, *Bmp4* expression is maintained within the embryonic shield that is relocated to the animal pole by the cell movements of gastrulation (Nikaido et al. 1997). *Bmp4* is first expressed in mesoderm at 10 hpf (Martinez-Barbera et al. 1997) and by 16 hpf *Bmp4* is expressed in lateral plate mesoderm (Dick et al. 1999) (Chocron et al. 2007). Since the deletion of ECR2 in mouse 5' GFP*lacZ*-BAC results in loss of *Bmp4* directed expression in mouse lateral plate mesoderm (Chapter IV) and *Bmp4* is also expressed in zebrafish lateral plate mesoderm, ECR2 may function as a lateral plate mesoderm enhancer in both fish and mouse.

Locating and studying an enhancer that is conserved in fish and mouse may provide insight into the impact evolution has had on the function of a cis-regulatory element. In addition, it provides evidence that *Bmp4* regulation in mesoderm is an ancient, ancestral vertebrate feature. Comparative analyses have shown that ECR2 is highly conserved between fish and mouse. Comparative analyses also indicate ECR2 is a unique sequence flanking *Bmp4* and is not conserved in any other locus of the genome including *Bmp4*'s close homolog, *Bmp2* (Chapter II). In Chapter V and VII, we provide evidence that ECR2 directs *Bmp4* expression in mesoderm using transgenic mouse models. Since ECR2 has been maintained in fish and mouse, we would hypothesize ECR2 has a similar function in fish. Whether or not the function of ECR2 has been maintained over 450 million years of evolution, understanding the function of this enhancer in both species would allow more detailed comparisons between sequences and, in turn, provide insight into the role transcription factor binding motifs play in the enhancer's function.

As discussed previously, deletion of ECR1 and ECR3 from 5' and 3' GFP*lacZ*-BACs did not clearly reveal a function for either ECR (Chapter V). These results still do not completely rule out a *cis*-regulatory role for ECR1 or ECR3 because the deletions may not have been large enough or the site of expression may be very discrete spatiotemporally and, therefore, missed during analysis of embryos at selected time points. Using zebrafish to test the enhancer activity of ECR1 and 3 may prove to be a relatively quick and inexpensive way to determine whether or not additional studies of these ECRs are worthwhile.

To test the enhancer activity of each ECR *in vivo*, a heterologous promoter reporter construct was coinjected with each ECR. Upon coinjection of linear DNA fragments in fish, random arrangements of head to tail and head to head concatemers are rapidly formed and replicated extrachromosomally (Iyengar et al. 1996) (Muller et al. 1997). Sometimes, the concatemer is randomly integrated into the genome (Stuart et al. 1988) (Bishop and Smith 1989) (Palmiter and Brinster 1986), while remaining exogenous DNA is degraded in post-gastrula stages (Stuart et al. 1988) (Iyengar et al. 1996). This method is convenient because 1) ECR sequences do not have to be cloned into the promoter reporter construct, and 2) others have shown this method enables efficient rates of transgenesis (Woolfe et al. 2005).

To test each fish/mouse ECR for enhancer activity in fish, microinjections were performed on fertilized zebrafish eggs and GFP expression was analyzed in transient transgenic fish. These experiments indicated ECR2 and ECR3 were capable of exhibiting enhancer activity in fish. ECR2 reliably directed GFP expression in the notochord at 24 hpf, while ECR3 directed expression in muscle at 24 hpf. ECR1 failed to exhibit enhancer activity in the transient transgenic fish assay.

Material and Methods

Identification of ECRs in Zebrafish Genome

To verify pufferfish ECRs were also present in the zebrafish genome, each pufferfish ECR sequence (Chapter II) was submitted to the BLAT search program on the UCSC Genome Browser (Zebrafish June 2004 Assembly).

Zebrafish Husbandry

Zebrafish AB strains were maintained according to conventional methods (Westerfield 2000). Embryos were generated by pairwise matings, maintained at 28.5° C, and staged according to approximate hours post fertilization (hpf). All studies were approved by the Vanderbilt University Institutional Animal Care and Use Committee.

DNA Constructs

A human βglobinGFPpA reporter construct was generously provided by Dr. Greg Elgar (Woolfe et al. 2005). Briefly, βglobinGFPpA plasmid DNA was purified and 2 ng were used as template in PCR reactions. βglobinGFPpA was amplified using the following primers: 5'-GGAAGGCCATCCAGCCTC-3' (forward) and 5'-GTGCCACCTGACGTCTAAG-3' (reverse). A high fidelity mixture of Taq and Pfu polymerases and an annealing temperature of 57.6°C was used to amplify the 1.5 kb construct. PCR reactions were purified using standard

phenol:chloroform procedures followed by ethanol precipitation. PCR reactions were verified by gel electrophoresis alongside lambda-HindIII digested DNA.

Zebrafish ECRs and control sequences as well as mouse ECRs were amplified from zebrafish genomic DNA or mouse BAC DNA using primers and annealing temperatures outlined in Table 6.1. Note, zebrafish ShhECR6 was identified by BLAT analysis using a 596 bp sequence of pufferfish ShhECR6 (Woolfe et al. 2005). A high fidelity blend of Taq and Pfu polymerases was used to decrease the incidence of mutations in PCR reactions. PCR-amplified ECRs and control sequences were purified using standard phenol:chloroform methods followed by ethanol precipitation. ECRs and control sequences were verified by gel electrophoresis alongside Lambda-HindIII digested DNA prior to microinjection.

Microinjections

ECR or control DNA was diluted to a final concentration of 225 ng/µl in solution with βglobinGFPpA at a final concentration of 25 ng/µl in 10 mM Tris. One microliter of 0.01% phenol red solution was added to the mixture immediately preceding injections to allow visualization of the DNA being injected. Fish embryos were collected from natural matings and injected at the 1 to 2-cell stage. At approximately 6 hpf, injected embryos were observed and any embryos that appeared abnormal or dead were discarded.

Analysis of Transient Transgenic Fish

Chorions were manually removed from injected embryos the following day and embryos were anesthetized using 3-aminobenzoic acid ethyl ester (Tricaine) and observed at 24 hpf for GFP-expressing cells under fluorescent illumination. Note, the total number of surviving injected fish was lower for the zebrafish ECR 1-3 constructs (see TABLE 6.2). This was due to a dramatic increase in the number of dead embryos that were culled at 6 hpf. The increase in embryo death in these particular injections was most likely due to the inadvertent use of small-bored pipets to transfer the embryos post-injection. Although the number of surviving injected fish was low, the percentage of GFP-positive fish was comparable to the positive control (Shh) (see TABLE 6.2).

Whole Mount *In Situ* Hybridization

Whole mount *in situ* hybridzation (WISH) was performed on 24 hpf zebrafish embryos using previously described methods as a guide (Jowett 2001). Embryos were fixed in 1x phosphate buffered saline (PBS) containing 4% paraformaldehyde overnight at 4°C then washed twice in 1x PBS containing 0.1% Tween 20 (PBT) at 4°C. Embryos were dehydrated through a graded series of methanol:PBT solutions and stored at -20°C for approximately two months. Prior to WISH, embryos were rehydrated through a graded series of methanol diluted in PBT. Rehydrated embryos were subjected to proteinase K (0.01 mg/mL final concentration) digestion for 15 minutes at room temperature and postfixed in 4% paraformaldehyde at room temperature for 20 minutes. Next, embryos were prehybridized in hybridization buffer (0.1% Tween, 1M citric

acid, 500 µg/ml yeast RNA, 5 X SSC, 50 µg/ml heparin, 50% formamide) at 68°C for 2 hours. Following prehybridization, embryos were hybridized overnight at 68°C with 0.3 µg/ml digoxygenin-labeled *Bmp4* RNA probe. The *Bmp4* probe template was generously provided by Dr. Lilianna Solnica-Krezel. The *Bmp4* probe was linearized with EcoRI and labeled with DIG-UTP using T7 polymerase (Roche, catalog# 1-175-025). Following posthybridization stringency washes, embryos were incubated in preabsorbed alkaline phosphatase conjugated anti-digoxygenin antibody (1:1000 in TSA block) for overnight at 4°C. Embryos were stained using the alkaline phosphatase substrate BM Purple (Roche, catalog #11442074001). Signal was observed after 2 hours of exposure to BM Purple.

Figure 6.1. Reporter construct used to test potential enhancer sequences for reporter activity. This construct was generously provided by Dr. Greg Elgar (Woolfe et al. 2005). Briefly, PCR primers (small arrows) were used to amplify the human βglobin heterologous promoter containing an EGFP reporter and polyadenlyation signal. The 1.5 kb amplified construct was coinjected with ECRs to assay enhancer function. Note, meganuclease sites (SceI) flank the promoter and have been shown to increase transgenic efficiency when contruct is injected as an intact, circular plasmid (Thermes et al. 2002). However, the meganuclease sites do not serve a purpose in our coinjection assays.

Table 6.1. Primers and annealing temperatures used to amplify ECR sequences.

Zebrafish ECRs	Forward primer	Reverse primer	Annealing temp (°C)
ECR1	TGCATAACTGAGCCAAACTGA	GCTGGATTGAGTCTGATCTGC	56.1
ECR2	GAAGCCGCGAGTACTGTGTT	CGAGCGTTAACCGTGTCTTT	56.8
ECR2_985bp	TATTGAAAATCGCGACCACA	TGAAAGCTCGGTGTCAACAG	55.0
ECR3	CTAAGCGGCCCTGACACTT	AAAAGTGCCGTTGTTGGAAG	56.8
Mouse ECRs			
ECR1	TTAATGGGCCACATCATCCT	CCAGAGACGGATGGCTAATG	57.6
ECR2	AACTGTGTCTCTTCAAAACTGA-CATT	CCTCTTCTCCCAGCCCTCT	58.2
ECR3	CCGGGCCACTTACAAATAAAA	GGAGGAACACAAAGA-TAAGGTCA	56.1
Controls			
SHH_6 (Woolfe et al. 2005)	CGAGCGGAGTTGGGATATT	GCATGTGCCTGTCCCACT	56.8
NCNC1	GAGAATGCAAAAGCATTGTTACAG	TGCTAAGCGCAATGTTTTGT	59.6
NCNC2	CGTGGCCTAAAGCTGATTGT	CACGTAGCGCTCAAGTAGCA	56.8
NCNC3	TGCTGGAGAACAGAGAAGCA	TCAGTTTAATTGATGCAAGTTTCC	55.1
Bmp4Exon4	ACTCATATCCACCGCAGAGC	GTTTTTCAGCACCACCCTGT	56.8

Results

Pufferfish ECRs Identify Zebrafish ECRs

Three ancient ECRs flanking *Bmp4* were identified by comparative analyses of pufferfish and mouse genomic sequences (see Chapter II). To test each sequence for enhancer activity, in vivo reporter assays must be employed in fish and/or mouse. To verify each ECR was present in zebrafish DNA, a BLAT search was performed on the UCSC Genome Browser using pufferfish ECR sequence against the zebrafish genome assembly (July 2007) (data not shown). Pufferfish ECR1 and 2 aligned to zebrafish chromosome 17, 5' to *Bmp4* (data not shown). Pufferfish ECR3 aligned to chromosome 17 as well and was located 3' to *Bmp4*, as expected (data not shown). Thus, comparative analysis suggests each pufferfish ECR is present in the zebrafish genome in the same order and orientation.

Zebrafish ECRs Exhibit Reporter Activity

To test the enhancer activity of each ECR *in vivo*, a human βglobinGFP promoter reporter construct was coinjected with each ECR. Once the zebrafish counterpart of each pufferfish ECR was identified (see above), zebrafish ECR 1-3 were amplified alongside mouse ECR 1-3. ECR sequences were coinjected with a human βglobin promoter EGFP reporter construct (Woolfe et al. 2005) (FIGURE 6.1). The promoter construct contained the human βglobin promoter followed by an EGFP reporter and a polyadenylation signal (FIGURE 6.1).

Coinjections were performed on at least 100 zebrafish embryos at the 1-2 cell stage for each construct tested. Injected embryos were screened at 6 hpf so that abnormal or arrested development could be culled. Later, at 24 hpf, GFP positive embryos were counted and the data was expressed as a percentage of total surviving embryos.

To control for microinjection technique and nonspecific reporter activity in the coinjection assay, control sequences were tested. More specifically, the previously published Shh enhancer (SHH_6) (Woolfe et al. 2005) was tested to verify good microinjection technique and results showed our technique (43% GFP-positive fish) was comparable to published results (44% GFP-positive fish, muscle-specific expression) (Woolfe et al. 2005) (TABLE 6.2). Next, the promoter construct was tested alone, resulting in a very limited number of GFP-positive fish (2.5% GFP-positive fish) (TABLE 6.2). Although the number of GFP-positive fish was higher than previously published results (0.5% GFP-positive fish) (Woolfe et al. 2005), it was significantly less than the least active element (ECR1, 33% GFP-positive fish) (TABLE 6.2). In addition, the few promoter construct-derived GFP-positive fish had extremely limited numbers of GFP-positive cells (TABLE 6.2). Taken together, this suggests the promoter contruct alone exhibited very little background activity. Finally, a randomly chosen noncoding, nonconserved sequence approximately 200 bp in length was coinjected with the promoter fragment, resulting in zero GFP-positive fish (TABLE

Table 6.2. Results from transient transgenic Zebrafish injections (dr= danio rerio, mm= mus musculus).

	GFP+ embryos/ Total embryos	% GFP+ embryos (# of injections)	GFP+ cells/ Embryo
Negative controls			
βglobinGFP	4/161	2.5% (2)	~1
NCNC-βglobinGFP	0/90	0% (1)	none
Positive control			
Shh_6-βglobinGFP	75/175	43% (1)	multiple
Experimentals			
drECR1-βglobinGFP	7/21	33% (1)	~2-4
drECR2-βglobinGFP	12/31	39% (1)	multiple
drECR2_985bp-βglobinGFP	26/40	65% (1)	multiple
mmECR2-βglobinGFP	186/249	75% (2)	multiple
drECR3-βglobinGFP	27/33	82% (1)	multiple

6.2). Thus, control DNA sequences failed to exhibit reproducible reporter activity (TABLE 6.2).

To test zebrafish ECR sequences for enhancer activity, each ECR was coinjected with the human βglobin promoter construct. Two out of three ECR sequences displayed specific, reproducible reporter activity in zebrafish at 24 hpf (TABLE 6.2). ECR1 displayed reporter activity (33% GFP-positive fish), however, the number of GFP-positive cells were very limited in individual fish (TABLE 6.2). Nevertheless, compared to the negative controls, ECR1 upregulated GFP expression (TABLE 6.2 and see FIGURE 6.2). ECR2 and ECR3 displayed more reproducible reporter activity, with multiple GFP-positive cells in individual fish (TABLE 6.2). The percentage of GFP-positive fish obtained with ECR2 and 3 was comparable to the positive control (SHH_6) (TABLE 6.2).

ECR2 Directs Expression in Mesodermally-Derived Notochord

As stated previously, ECR2 and ECR3 displayed reproducible reporter activity in zebrafish at 24 hpf (see TABLE 6.2 and FIGURE 6.2). ECR2 directed specific GFP expression in multiple cells along the midline of the fish (FIGURE 6.3). The ECR2-directed GFP-positive cells were consistent in their size and shape (FIGURE 6.3). Close examination revealed ECR2-directed GFP expression specific to notochord cells (FIGURE 6.4). Interestingly, both the zebrafish and mouse ECR2 sequence reproducibly directed GFP expression in notochord cells (see TABLE 6.2).

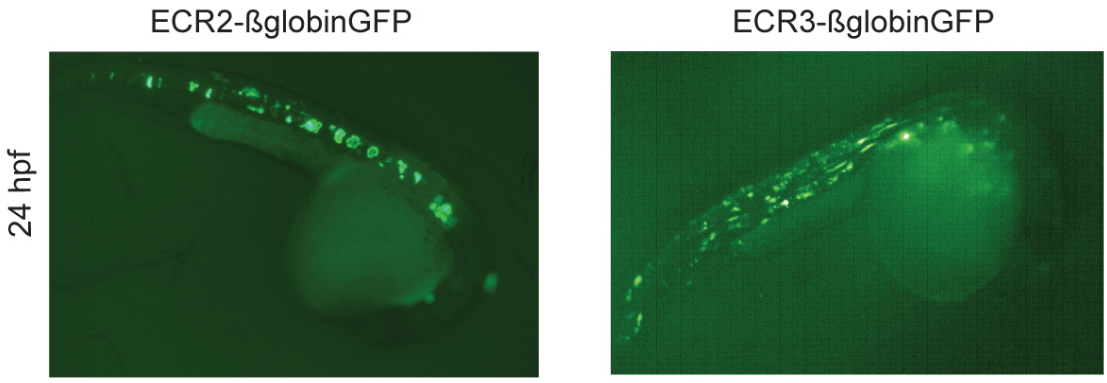

Figure 6.2. ECR sequences direct reporter activity in zebrafish at 24 hpf. Zebrafish ECR2 directs expression in multiple notochord cells while ECR3 directs expression in multiple muscle cells as well as an undetermined population of cells. Note, expression is mosaic with the transient transgenic assay.

ECR3 directed GFP expression in numerous cells dispersed throughout the trunk and tail of the fish (FIGURE 6.2). A subset of GFP-positive cells were narrow and oblong, while the other population of cells was more compact and rounded (FIGURE 6.2). The narrow, oblong cell population was muscle and these cells were concentrated in the trunk region (FIGURE 6.2). The other population of ECR3-directed GFP expression in rounded cells were most likely muscle precursor cells or myoblasts. These cells were also distributed throughout the trunk, alongside the muscle cells.

Although ECR2 specifically directed GFP expression in notochord cells at 24 hpf, there was no evidence in the literature that *Bmp4* was endogenously expressed there. Therefore, we performed *in situ* hybridization to determine whether or not *Bmp4* was expressed in zebrafish notochord cells at 24 hpf. In situ hybridization confirmed *Bmp4* is not expressed in notochord cells at 24 hpf (FIGURE 6.3). Likewise, *Bmp4* was not detected in muscle cells in contrast to ECR3-βglobinGFP coinjections. *Bmp4* expression was detected in the intermediate cell mass of mesoderm (FIGURE 6.3, arrow) as previously reported (Leung et al. 2005), as well as dorsal retina, otic vesicle, heart, and nose (data not shown) (Thisse et al. 2004).

Since ECR2 is conserved in mouse and mouse ECR2 also directed expression in zebrafish notochord (TABLE 6.2), we wanted to explore the possibility that *Bmp4* may be expressed in mouse notochord. To look at *Bmp4* expression in mouse notochord, *Bmp4* $^{lacZ+/-}$ embryos were generated at 9.5 dpc and stained with Xgal to detect *lacZ* expression (FIGURE 6.4). Embryos were

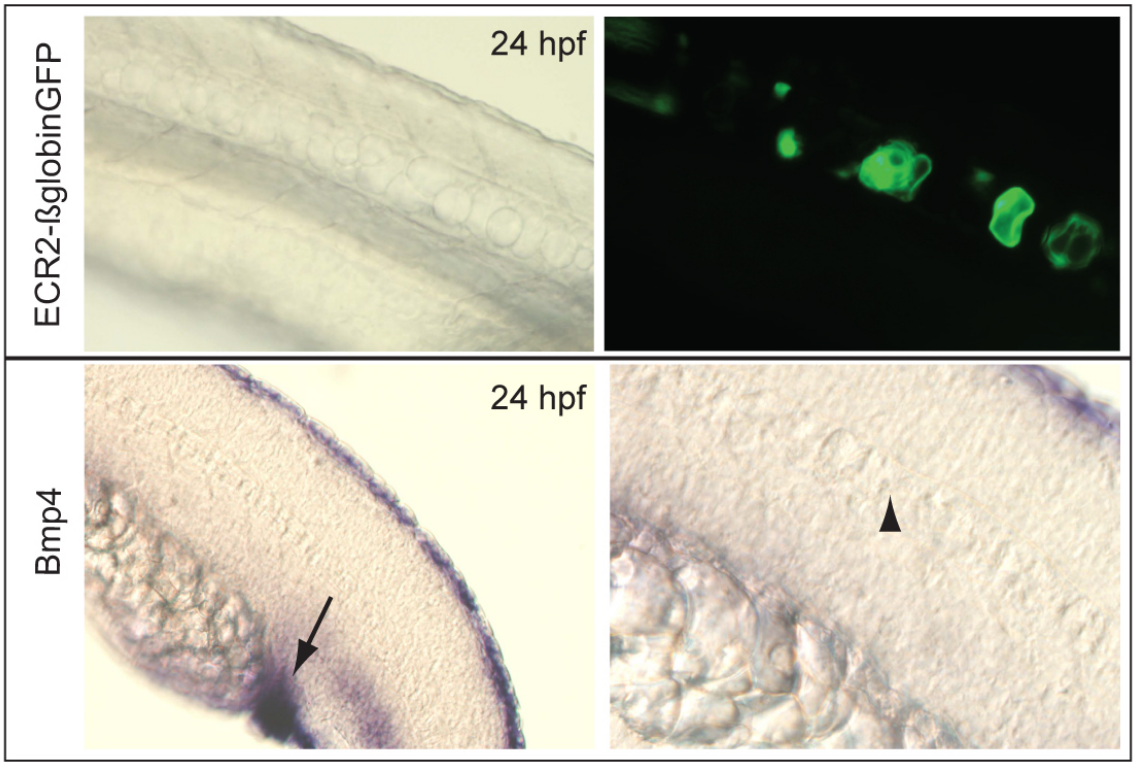

Figure 6.3. ECR2 directs expression in notochord. (Top panels) A high magnification brightfield image of a transgenic ECR2-βglobinGFP fish alongside an image of the same region under UV light clearly shows GFP fluorescence is located in the notochord at 24 hpf. (Bottom panels) In situ hybridzation performed on 24 hpf wild type zebrafish using a zebrafish Bmp4 probe reveals Bmp4 *is not expr*essed in notochord (black arrowheads) at 24 hpf. Expression was observed in other reporter sites of *Bmp4* expression such as the intermediate cell mass of mesoderm (arrows).

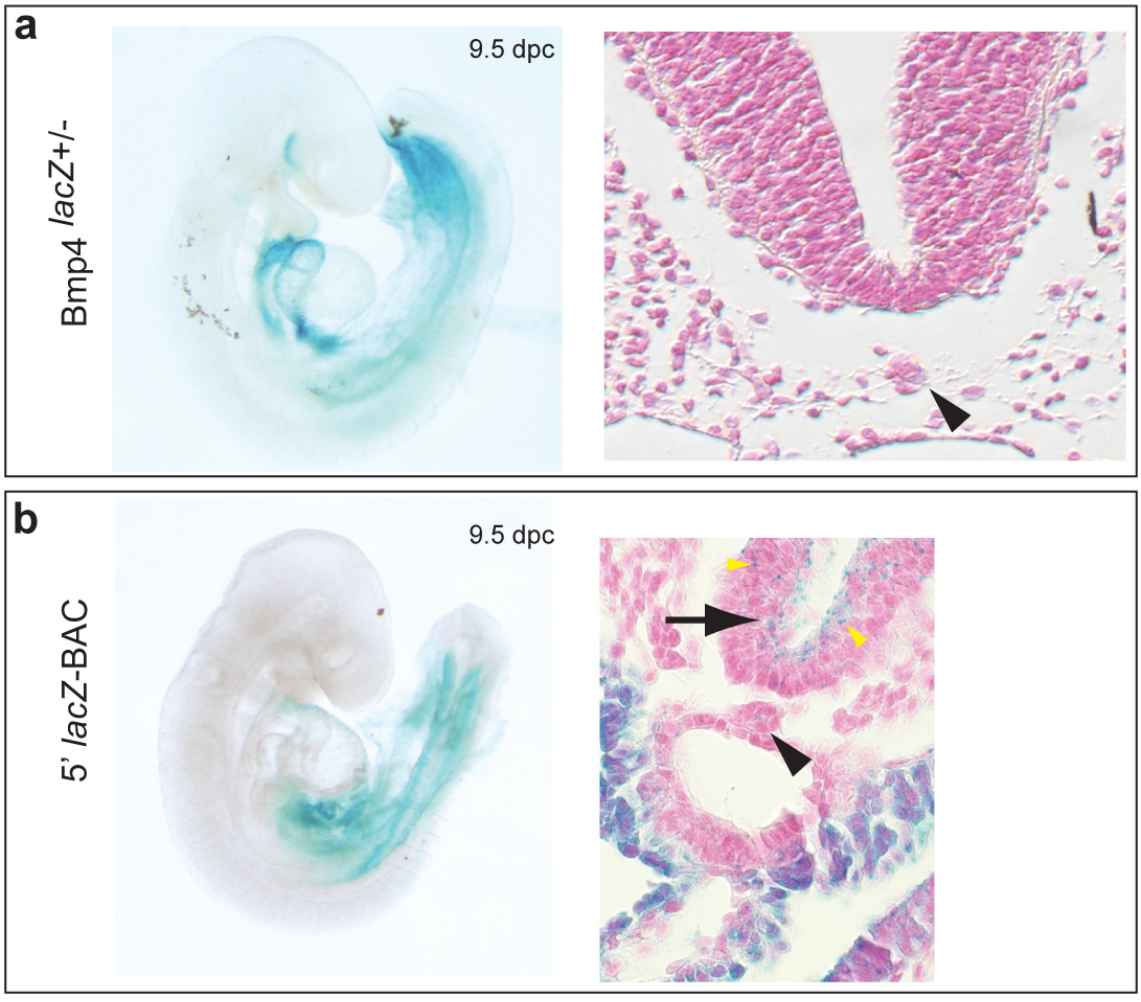

Figure 6.4. *Bmp4* is not expressed in mouse notochord. (a) Whole mount Xgal staining to detect *lacZ* expression in *Bmp4* $^{lacZ+/-}$ embryos indicates endogenous *Bmp4* expression at 9.5 dpc. Serial sections through 9.5 dpc embryos shows *Bmp4* is not expressed in mouse notochord (arrowhead) as indicated by a lack of blue staining. (b) Whole mount Xgal staining to detect *lacZ* expression in 5' GFP*lacZ*-BAC embryos at 9.5 dpc. Histological sections revealed sporadic, punctate staining in notochord (arrowhead) and neural tube (arrow) in 9.5 dpc 5' GFP*lacZ*-BAC embryos.

serially sectioned allowing a complete view of the notochord along the anterior-posterior axis. Analysis of these sections failed to detect *lacZ* expression in mouse notochord at 9.5 dpc (FIGURE 6.4a, arrowhead). ECR2 is present in the 5' GFP*lacZ*-BAC, therefore, embryos from 5' BAC lines were generated at 9.5 dpc and stained with Xgal to detect *lacZ* expression as well (FIGURE 6.4b). Upon close inspection of histological sections through the notochord of 5' GFP*lacZ*-BAC embryos, we observed sporadic, punctate staining in the notochord (FIGURE 6.4b, arrowhead). Staining was only seen in one or two cells per section and never in every cell comprising the notochord. Interestingly, staining was not observed throughout the cytoplasm of the cell as would be expected in this line as the β-galactosidase is cytoplasmic. In addition, a similar punctate staining pattern was observed in the neural tube of 5' GFP*lacZ*-BAC embryos but not in *Bmp4* $^{lacZ+/-}$ embryos (FIGURE 6.4b, arrows). The punctate staining present in 5' BAC neural tube and notochord may be persisting β-galactosidase in vacuoles targeted for degradation because *lacZ* is expressed much earlier in development in mesodermal precursors.

ECR2 Fails to Direct Expression in Early Mesoderm in Fish

As discussed in Chapter III, initial analysis of 5' GFP*lacZ*-BAC, but not 3' GFP*lacZ*-BAC mouse embryos, at 9.5 dpc revealed *lacZ* expression in lateral plate mesoderm (see FIGURE 3.10). In addition, analysis of embryos earlier in development showed the 5' GFP*lacZ*-BAC directed expression in extraembryonic mesoderm (see FIGURE 3.10). Likewise, deletion of ECR2 from the 5'

GFP*lacZ*-BAC resulted in the loss of mesodermal expression (see Chapter V). In zebrafish, the notochord is mesodermally-derived from the embryonic shield (Stemple 2005). *Bmp4* is expressed in the inner cells of the embryonic shield at 6 hpf (Nikaido et al. 1997) (Wang et al. 1999) and at 10 hpf in the prechordal plate (Solnica-Krezel and Driever 2001) which arises from the shield. Therefore, to investigate the possibility that ECR2 may direct expression in mesodermally-derived tissues that precede notochord development in fish, we analyzed ECR2-βglobinGFP injected fish at 6 and 10 hpf for GFP activity. Although GFP activity was detected in notochord at 24hpf, no GFP activity was detected earlier in development (6 and 10hpf), when *Bmp4* is expressed in early mesodermal cells (data not shown).

Initially, ECR2 was amplified as a 249 bp segment identified by BLAST analysis using pufferfish ECR2. To test for critical sequences beyond the initial 249 bp tested, we also amplified a 985 bp sequence containing ECR2 (TABLE 6.1 and FIGURE 6.5) and coinjected this larger sequence with βglobinGFP. The 985 bp sequence was chosen because it was the largest amount of sequence containing ECR2 that could be readily amplified since large repetitive stretches were present on either side of this interval. In addition, the 985 bp sequence extends well beyond the sequence conservation on both sides (FIGURE 6.5). Injected fish were observed at 6, 10, and 24 hpf to look for expression in early mesoderm tissues (embryonic shield, prechordal plate) and the mesodermally-derived notochord. Like the smaller ECR2 fragment tested, the larger fragment directed notochord-specific expression at 24 hpf. However, the larger fragment

failed to direct GFP expression at 6 and 10 hpf (TABLE 6.2 and data not shown). Taken together, ECR2 directs GFP expression in zebrafish notochord at 24 hpf, but not in mesoderm at 6 or 10 hpf.

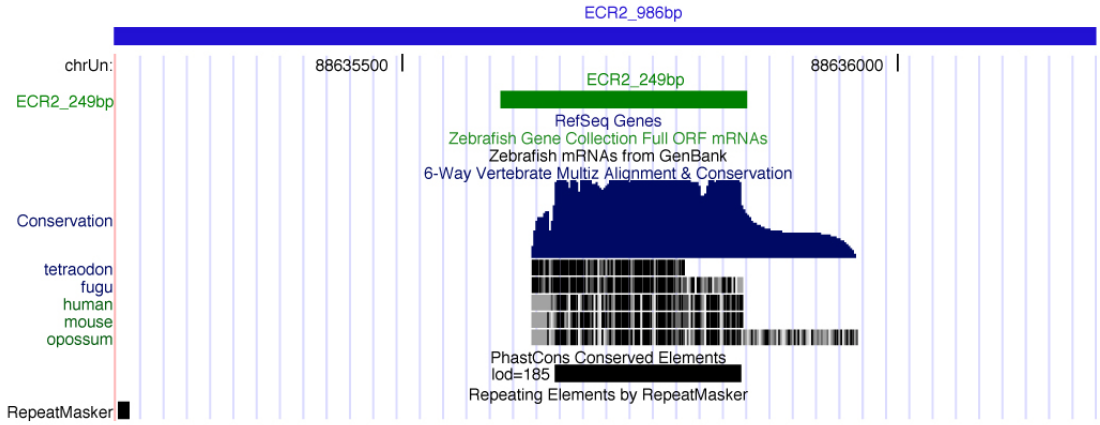

Figure 6.5. Custom tracks on the UCSC Genome browser June 2004 genome assembly (http://genome.ucsc.edu/) depict two ECR2 fragments tested in zebrafish. The original ECR2 fragment tested was 249 bp and is indicated by the green bar. A 985 bp fragment containing the orginal 249 bp ECR2 is indicated by the blue bar. Large stretches of repetitive sequence flank the 985 bp fragment as indicated by the "Repeating Elements masked by RepeatMasker" track (black bar). Note, the 3' repetitive stretch is not shown.

Discussion

Although numerous diseases have been shown to result from coding mutations, many others do not have coding mutations and are thought to be affected by noncoding mutations that impact gene regulation. In addition, little is known about the function of the noncoding genomic landscape. Therefore, it has become increasingly apparent that the location, annotation, and functional assessment of noncoding elements as well as the development of high-throughput assays to perform these tasks are a priority. To this end, the genomic landscape encompassing *Bmp4* is an excellent candidate for addressing the function of noncoding elements because it resides in a gene desert and contains a significant amount of noncoding conservation (Chapter II). In addition, the location of three noncoding ECRs flanking *Bmp4* that are present in mouse and pufferfish allows us to test the function of each ECR using a high-throughput transient transgenic fish assay.

Using the sequence of each pufferfish ECR, we identified all three orthologous ECRs in the zebrafish genome. In addition, BLAT analysis of each ECR using the July 2007 assembly on the UCSC Genome Browser reveals each zebrafish ECR is located in the same order and orientation, with respect to *Bmp4*, as they are in other vertebrates (see FIGURE 2.5 and data not shown). The confirmation of each ECRs existence in the zebrafish genome allowed us to test zebrafish ECR sequence rather than pufferfish sequence in the zebrafish transient transgenic assay.

Performing coinjections of ECR sequences with a heterologous promoter in transient transgenic zebrafish assays allowed us to quickly test mutiple sequences for reporter activity *in vivo*. Two of three ECRs exhibited highly reproducible, tissue-specific reporter activity *in vivo*. ECR1 exhibited reporter activity that was difficult to interpret tissue-specificity due to the small number of GFP-positive cells and limited number of embryos. Our inability to detect reproducible expression may be due to the mosaic nature of expression when working with this transient transgenic assay. In zebrafish transient transgenic assays, reporter expression is highly mosaic, resulting in, at best, a portion of the cells in a structure actually expressing the transgene. Since ECR1 directed transgene expression in a limited number of cells (Table 6.2) per fish, it was more difficult to determine whether or not the expression was tissue-specific. Therefore, additional injections of ECR1 may shed light on its tissue-specificity. In particular, if the tissue that ECR1 directs expression in is a small population of cells to begin with or is highly mosaic then the results may be difficult to interpret. Stable transgenic lines can be generated from injected fish that have incorporated the transgene into the genome and are able to transmit the transgene through the germline. Unlike transient transgenic fish, stable transgenic fish lines can exhibit non-mosaic reporter expression, allowing the enhancer-driven expression to be potentially present in every cell of the structure of interest. The drawbacks to this method are that it is substantially slower than transient transgenic methods and germline transmission is not always efficient. However, it would be beneficial to generate stable transgenic lines for these

constructs and analyze them for tissue-specific expression as well as to compare the results to the transient transgenic method.

Tissue-specific reporter activity was observed with ECR2 and ECR3. ECR2 directed expression in notochord, while ECR3 directed expression predominately in what appears to be muscle cells. However, *Bmp4* was not endogenously expressed in either of these tissues as verified by in situ hybridization. Therefore, we cannot rule out the possibility that while the assay is able to test the enhancer activity of noncoding ECRs, it is not capable of elucidating tissue-specificity of the noncoding ECRs we tested. For example, ECR2 directed expression reproducibly in mesodermally-derived notochord tissue. Perhaps the actual function of ECR2 in zebrafish is to direct expression specifically in the mesodermal tissue that precedes notochord development, such as in the embryonic shield. Although ECR2 did not direct expression in the embryonic shield early in fish development, it is possible that the transient transgenic assay isn't capable of properly assaying transgene activity during early fish embryogenesis. It is possible that the time it takes to transcribe and translate GFP allows a large window of early and rapid fish development to pass by before the transgene can be detected. In support of this, Stuart *et al.* reported that subsequent to injection of DNA sequences in fish embryos, transgene sequences were not detected until 6-10 hpf and the majority of these sequences degraded during gastrulation (Stuart et al. 1988). Therefore, understanding the tissue-specificity of ECR sequences that act as enhancers early in development may be very difficult to glean from transient transgenic assays. As stated

previously, it may be necessary to generate stable transgenic lines to assay the function of ECRs in fish.

Interestingly, 5' GFP*lacZ*-BAC mouse embryos exhibit subtle staining patterns in the mouse notochord at 9.5 dpc (see FIGURE 6.5). Although *Bmp4* is not normally expressed in the notochord of fish or mouse, it is possible that a repressor is present beyond the confines of the 5' BAC interval that normally represses the notochord expression activated by ECR2. In that case, if ECR2 was removed from the endogenous locus, it should be able to activate transcription in the notochord because it has been removed from the context of the repressor. Analysis of ECR2 minigenes in transgenic mice would be required to corroborate this hypothesis (Chapter VII).

Alternatively, it is possible that the ECRs identified actually regulate an adjacent gene rather than *Bmp4*. However, comparative analysis suggests the ECRs have been maintained in a syntenic block with *Bmp4* over millions of years of evolution, while the flanking genes have not, lending weight to the idea that these ECRs are important for *Bmp4* regulation (see Chapter II). However, the ECRs tested may require the endogenous Bmp4 promoter to function. More specifically, the endogenous promoter may contain sequences that are required for interactions that allow the ECR to upregulate Bmp4 expression. As mentioned previously, we have identified a zebrafish BAC clone that contains the *Bmp4* promoter and could be used to amplify and clone the promoter into a reporter vector (see Chapter II). This would allow future studies to test each ECR in the context of the endogenous promoter.

Although the coinjection assay is a quick and efficient method for producing multiple transgenic fish, it may not be the best method to assay enhancer function. Alternative strategies could be used to compare high throughput techniques that would allow numerous sequences to be rapidly tested in vivo. For example, ECRs could be cloned into the promoter construct presented here (FIGURE 6.1) and injected as a circular plasmid with meganuclease to increase transgenic efficiency and decrease mosaicism in founder fish (Thermes et al. 2002). Coinjection of a transgene flanked by SceI sites (FIGURE 6.1) with meganuclease enzyme has been shown to increase promoter-based expression and dramatically increase germline transmission rate to allow the generation of stable lines. Another alternative method takes advantage of the fish-specific transposon Tol2 (Kawakami et al. 2004). Transgenic expression appears more consistent and rates of germline transgenesis are higher using this method (Allende et al. 2006). Taken together, there are multiple alternative options to the coinjection methods presented here.

In sum, transient transgenic fish assays provided an efficient way to test multiple DNA sequences for enhancer-like activity. Although the precise function of each ECR could not be determined by this method, it may be a cost effective screening tool to test a large number of sequences *in vivo* for reporter activity. This is especially important for sequences of unknown function since it would be impossible to know what type of cell line could be used to assay reporter activity.

CHAPTER VII

ECR2 IS SUFFICIENT TO DIRECT MESODERM EXPRESSION IN MOUSE

Introduction

Increasing evidence throughout the previous Chapters of this thesis suggest a modular enhancer element resides nearly 50 kb 5' to mouse *Bmp4*. More specifically, Chapter II showed a highly conserved, ancient DNA element with multiple transcription factor binding motifs located in the same orientation relative to *Bmp4* in human, mouse, and pufferfish. In addition, Chapter III demonstrated that a 5' GFP*lacZ*-BAC transgene is able to direct expression in mesoderm, while a 3' GFP*lacZ*-BAC transgene fails to direct mesoderm expression suggesting a mesoderm enhancer is located in a –28 to -199 kb interval within the 5' BAC. Chapter IV showed a seamless deletion of mouse ECR2 from the 5' GFP*lacZ*-BAC results in complete tissue-specific loss of *lacZ* expression in mesoderm. Finally, transient transgenic fish assays indicated ECR2 exhibits enhancer activity *in vivo*. Taken together, these data suggests ECR2 is an ancient, long-range enhancer required in mammals for mesoderm expression. However, it does not address the question of whether ECR2 is sufficient for mesoderm expression. Therefore, this Chapter addresses the ability of ECR2 alone to direct *Bmp4* expression in mouse mesoderm via a heterologous promoter.

In addition to ECR2, two other ancient noncoding conserved DNA sequences were identified by comparative analysis adjacent to *Bmp4* (Chapter

II). Although deletion experiments in Chapter V failed to demonstrate a tissue-specific role for ECR1 and ECR3, transient transgenic experiments in fish (Chapter VI) suggested that ECR 3 may function as a tissue-specific enhancer. Therefore, this Chapter addresses whether ECR1 and ECR3 are sufficient to direct tissue specific expression in transgenic mouse assays.

To test the sufficiency of ECRs for enhancer function, the same heterologous promoter used in transient transgenic fish assays (βglobin)(Chapter VI) was utilized in stable and transient transgenic mouse experiments. In addition, ECR2 was tested with an alternative heterologous promoter, Hsp68. While ECR1 and ECR3 failed to show enhancer function in mice, ECR2-βglobin*lacZ* transgenic mice allowed us to define an ancient mesoderm enhancer located approximately 50 kb from the *Bmp4* promoter.

Material and Methods

ECR-βglobin*lacZ* and ECR2-Hsp68*lacZ* Constructs

To generate ECR-βglobin*lacZ* constructs, mouse ECR sequences were amplified using Expand High Fidelity Plus PCR kit (Roche). ECR1 (207 bp) and 2 (220 bp) were amplified using 5' GFP*lacZ*-BAC DNA as a template, while ECR3 (179 bp) was amplified using 3' GFP*lacZ*-BAC DNA as a template. Primer sequences used were previously outlined in Table 6.1. For ECR2, the following primers were used to amplify increasingly larger sequences containing the core ECR2 (220 bp) sequence: for ECR2-467 bp, (forward)

GAGTCTCCTTTCAGCCTTGC; (reverse) CCCTTCTGGGGATGAAAGTA and for ECR2-668 bp, (forward) TTCCACTTTGCTTCCCAAAC; (reverse) GGGGATGAAAGTAGCATCCTG.

Next, ECRs were ligated into pGEM-Teasy using the pGEM-Teasy Vector System I kit (Promega). Restriction digests were performed with NotI enzyme to isolate each ECR from pGEM-Teasy and NotI ECR fragments were subcloned into pBGZ40 (βglobin/*lacZ*) (Maconochie et al. 1997) and pSfi-Hsp68/*lacZ* in the forward orientation. ECR plasmids were verified by direct sequencing. After verification, the following clones were selected for purification and pronuclear injection: ECR1-βglobin/*lacZ* (Clone 5), ECR2-βglobin/*lacZ* (Clone 5), ECR2-467bp-βglobin/*lacZ* (Clone 5), ECR2-668bp-βglobin/*lacZ* (Clone 4), ECR3-βglobin/*lacZ* (Clone 1), ECR2-467bp- Hsp68/*lacZ* (Clone 2), ECR2-668bp-Hsp68/*lacZ* (Clone 2).

Purification of Plasmid Transgenes for Pronuclear Injection

ECR-βglobin/*lacZ* plasmids were digested with XhoI, XmnI, and SacII or XhoI and NgoMIV to isolate ECR-βglobin/*lacZ* from the vector backbone. ECR-Hsp68/*lacZ* plasmids were digested with XhoI and NgoMIV to isolate ECR-Hsp68/*lacZ* from the vector backbone. Digests were gel purified overnight using a low melting point agarose gel. Bands corresponding to the transgene fragment were excised and agarose was removed from the DNA with GELase (Epicentre® Biotechnology). After the agarose was digested with GELase, three phenol extractions were performed followed by one

phenol:chloroform:isoamyl alcohol extraction. Next, a chloroform extraction was performed and the DNA was then ethanol precipitated. Recovered DNA pellets were washed with 70% ethanol, resuspended in TE, then reprecipitated. The resulting DNA pellets were resuspended in microinjection buffer (10 mM Tris-HCL [pH 7.5], 0.15 mM EDTA [pH 8.0] in embryo grade water). DNA concentration was estimated by UV spectroscopy and DNA quality was assessed by gel electrophoresis.

Generation of Transgenic Mice

ECR-βglobin/*lacZ* and ECR-Hsp68/*lacZ* plasmids were submitted to the Vanderbilt Transgenic Core Facility for pronuclear injections of C57BL/6J x DBA/2J F1 hybrid embryos. DNA samples from yolk sacs or tail biopsies were used to verify transgenic embryos or weanlings by PCR methods.

Xgal Staining, Histology, Microscopy, and Imaging

Xgal staining, histology, microscopy and imaging were performed as described previously (Chapter 3).

Multi-Sequence Alignment and Binding Motif Identification

To perform an alignment of multiple sequences from multiple species, the Mulan tool was used from the DCODE suite of comparative analysis tools (http://mulan.dcode.org/) (Ovcharenko et al. 2005a). Sequences containing ECR2 were obtained from pufferfish, zebrafish, chicken and human by performing a BLAT analysis with ECR2 sequence (Chapter II) on the UCSC Genome Browser. Then, the aligned sequence and a large amount of sequence flanking the

alignment was obtained from the UCSC Genome Browser (Total sequence used: pufferfish, 1 kb; zebrafish, 1.5 kb; chicken, 1.4 kb; and human, 1.4 kb). Finally, the largest mouse sequence containing ECR2 (668 bp) was aligned with the pufferfish, zebrafish, chicken and human sequences.

To find predicted transcription factor binding sites in conserved sequences, the weight matrix-based MATCH™ tool from the TRANSFAC® database of transcription factors was utilized (Kel et al. 2003) (Matys et al. 2006). Unless otherwise noted, the profile was "vertebrate non-redundant minFP" and the cutoff selection for the profile used was "minimize false positives" (minFP).

Results

Initial ECR-βglobin*lacZ* Transgenes Fail to Direct Reproducible Reporter Expression in Mid-Gestation Mouse Embryos

As outlined in Chapter II, three ancient ECRs flank *Bmp4.* In addition, transient transgenic analysis of each ECR in zebrafish demonstrated enhancer activity for ECR 1, 2 and 3 (Chapter VI). To test the sufficiency of each ECR to direct a *Bmp4* expression pattern in mouse, ECRs were cloned in front of the minimal human βglobin promoter driving a *lacZ* reporter gene and subjected to pronuclear injection. This promoter was chosen because it was also used to test the enhancer activity of each ECR in zebrafish (Chapter VI) and it has been shown to be an effective heterologous promoter in mouse (Summerbell et al. 2000). ECR fragments were defined by pufferfish/mouse conservation (Chapter II). For each ECR, the pufferfish/mouse conserved segment as well as additional

flanking sequence (ECR1, 101 bp flanking sequence; ECR2, 86 bp flanking sequence; ECR3, 79 bp flanking sequence) (FIGURE 2.4) was initially tested.

Initial pronuclear injections were used to obtain founder mice, whereby breeding lines were established. A total of 19 lines were established from the three ECRs. Each line was assayed for reporter expression at three embryonic stages (9.5, 12.5, 15.5 dpc). Six lines were established for ECR1-βglobin/*lacZ* and four lines were established for ECR3-βglobin/*lacZ*. Xgal staining of transgenic embryos from all lines at each time point failed to demonstrate reproducible expression patterns (data not shown). In addition, there were no patterns of expression that recapitulated *Bmp4* expression (data not shown). Finally, nine lines were established for ECR2-βglobin/*lacZ* (L20, L21, L34, L63, L70, L85, L98, L104, L105). Since ECR2 was suspected to be a mesoderm enhancer and the 5' GFP/*lacZ*-BAC transgene was sufficient to direct extraembryonic mesoderm expression at 7.5 dpc in addition to lateral plate mesoderm expression at 9.5 dpc, ECR2-βglobin/*lacZ* transgenic lines were also assayed for reporter expression at 7.5 dpc. Like ECR1, none of the lines directed reproducible expression patterns (data not shown). In fact, 0/9 ECR2-βglobin/*lacZ* transgenic lines direct expression in extraembryonic mesoderm or lateral plate mesoderm.

Since initial tests of ECR-βglobin/*lacZ* constructs failed to exhibit reproducible *Bmp4* expression patterns, additional contructs were designed to test larger fragments, but only for ECR2. Because deletion of ECR2 from the 5'

GFP*lacZ*-BAC transgene resulted in the loss of *Bmp4* directed expression in mesoderm, our efforts focused solely on additional tests of the ECR2 region.

Although the initial ECR2-βglobin*lacZ* constructs failed to direct mesoderm expression, the 220 bp ECR2 fragment that was tested incorporated only the minimal, core pufferfish/mouse conserved sequence (FIGURE 7.1, green bar and see FIGURE 2.4). However, additional vertebrate multispecies conservation exists beyond the minimal 220 bp fragment as indicated by the black peaks at the bottom of the UCSC Genome Browser figure (FIGURE 7.1). Interestingly, there are two predicted conserved elements shown in the PhastCons Conserved Elements track (FIGURE 7.1). Each predicted conserved element is assigned a maximum likelihood of the odds ("lod") score of 12 and 139, respectively (FIGURE 7.1). Conserved elements in this track are predicted using a two-state phylogenetic hidden Markov model (phylo-HMM) (Siepel et al. 2005). The higher the lod score, the less chance the observed sequence identity could be due to chance under the neutral evolution model of the rate of mutation accumulation. Basically, higher lod scores tend to predict higher sequence conservation. The 668 bp ECR2 fragment encompasses both predicted conserved elements, while the 467 bp ECR2 fragment contains the predicted conserved element with the highest lod score (FIGURE 7.1).

To assess the level and extent of conservation in our region of interest at the base pair level, the Mulan tool was utilized (Ovcharenko et al. 2005a). Mulan allows multiple sequences from multiple species to be aligned and visualized in both a text and graphical format. Using the 668 bp mouse sequence containing

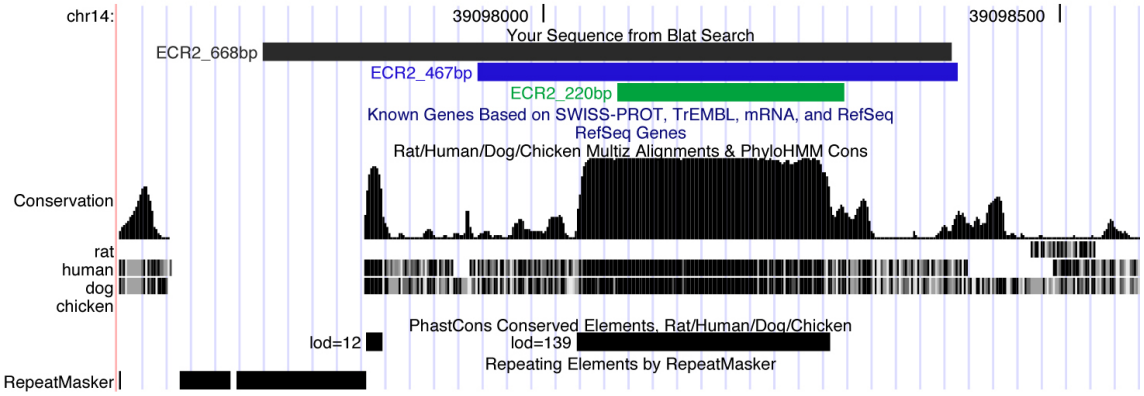

Figure 7.1. Distinct ECR2 fragments have increasing amounts of multi-vertebrate conservation. The UCSC Genome Browser (May 2004 Assembly) shows a segment of mouse chromosome 14 located approximately 50 kb 5' to *Bmp4* where ECR2 resides. Three ECR2 fragments were PCR amplified and tested *in vivo* for enhancer activity. The smallest ECR2 fragment was 220 bp (green bar) and contains most of the large, black peak of multi-vertebrate conservation as depicted by the Rat/Human/Dog/Chicken Multiz Alignment & PhyloHMM Cons track. The 467 bp ECR2 fragment (blue bar) contained the entire main peak of conservation and additional flanking sequence with minimal conservation. This fragment encompassed the entire PhastCons Conserved Element with a lod score of 139 indicating a predicted conserved element that is more likely conserved than nonconserved (Siepel et al. 2005). The largest ECR2 fragment was 668 bp (black bar) and included an adjacent smaller peak of conservation with a lod score of 112 as well as the main peak of conservation.

the core 220 bp ECR2 segment as a reference sequence (FIGURE 7.1a), Mulan was used to align pufferfish, zebrafish, chicken and human sequences (FIGURE 7.2a). In the resulting graphical visualization, the extent of conservation between mouse and the other species can be seen (FIGURE 7.2a). For example, the length of conserved sequence in chicken is longer than pufferfish or zebrafish, but shorter than human as indicated by the multicolored bars floating above each graphical plot (FIGURE 7.2a). Alignment of the ancient conserved sequence from each species shows the extent of conservation between mouse and each individual species down to the base pair level (FIGURE 7.2a and b). Multiple sequence alignment shows extensive conservation amongst vertebrates in the ancient conserved segment of ECR2 (FIGURE 7.2a and b). To account for conservation flanking the ancient conserved segment, additional ECR2-βglobin*lacZ* constructs were designed to incorporate the entire main peak of multispecies conserved sequence (FIGURE 7.1, 467 bp, blue bar) or the main peak as well as a short, highly conserved nearby peak (FIGURE 7.1, 668 bp, black bar).

Larger ECR2 Sequences are Sufficient to Direct Mesoderm Expression in Mouse

Since Deletion 2 BAC embryos suggested ECR2 was required for mesoderm expression and the 220 bp ECR2 sequence failed to direct mesoderm expression in embryos, we wanted to quickly test larger ECR2 sequences for the ability to direct mesoderm expression. Therefore, transient transgenic mouse embryos were generated by pronuclear injection. Five transgenic 8.5 dpc

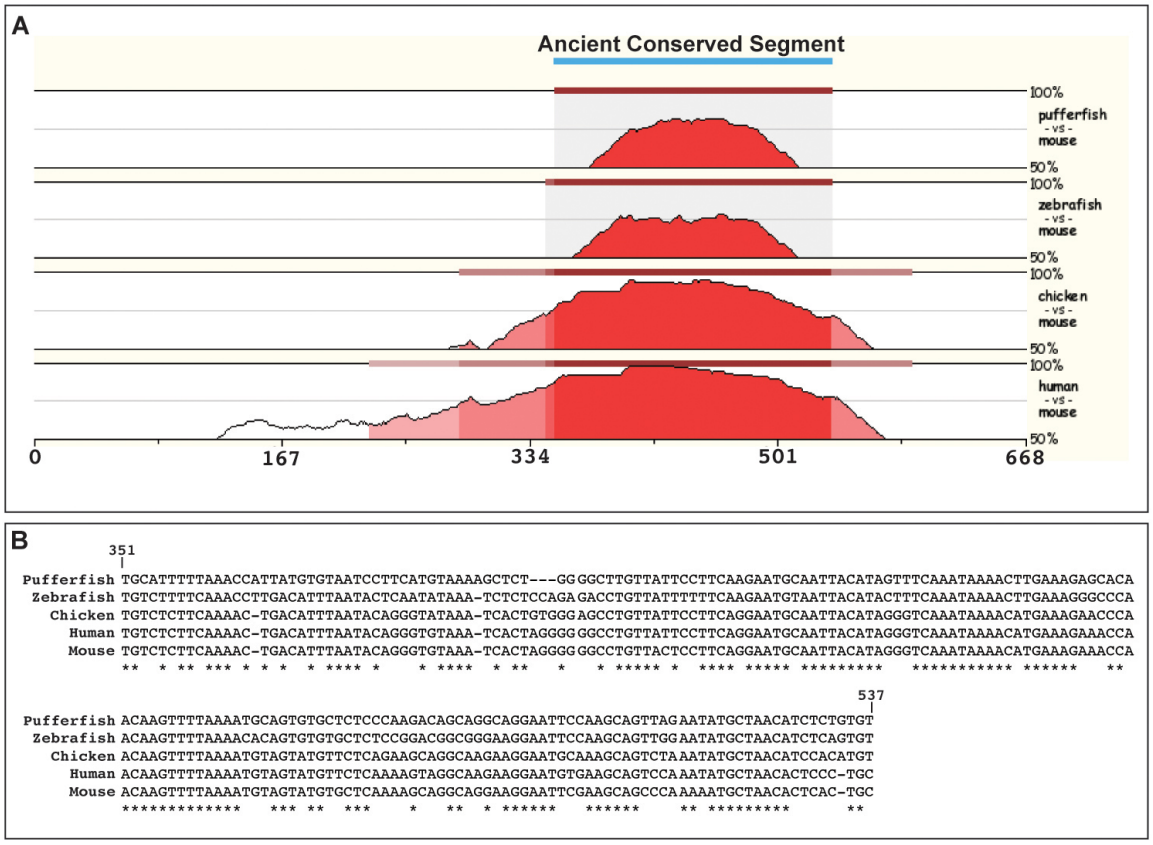

Figure 7.2. Multi-sequence alignments generated by Mulan depicted the conservation of ECR2 amongst mouse, human, chicken, zebrafish and pufferfish. (a) Graphical representation of multi-sequence alignments showed the interspecies conservation by color density. The dark red indicates conservation of sequence in all species tested. As the color density decreases, so does the number of species where the sequence is conserved. The 668 bp mouse ECR2 fragment was the reference sequence and it was aligned with the syntenic region in (from top to bottom) pufferfish, zebrafish, chicken and human. Note, the 467 bp mouse ECR2 fragment begins at base pair 209 of the 668 bp mouse ECR2 fragment and contains the remainder of the 668 bp fragment. (b) Mulan-generated sequence alignment of the ancient conserved segment depicts conservation at the base pair level.

embryos were generated from a 668 bp ECR2-βglobin/*acZ* construct. The 668 bp ECR2-βglobin/*acZ* transgene was sufficient to direct lateral plate mesoderm expression in 5/5 independent embryos (FIGURE 7.3a). This expression closely recapitulated the lateral plate mesoderm expression directed by the 5' GFP/*acZ*-BAC transgene that contained ECR2 (FIGURE 7.3a). Interestingly, the 467 bp ECR2-βglobin/*acZ* transgene was also able to direct lateral plate mesoderm expression in 3/3 8.5 dpc embryos (FIGURE 7.3a) suggesting the short, significant peak of conservation (FIGURE 7.1) is not required for mesoderm enhancer function. Note, the 467 bp ECR2-βglobin/*acZ* transgene incorporates the entire mouse/human ECR as depicted by the red shaded area beneath the curve in the mouse/human alignment (FIGURE 7.2a). Since the 5' GFPl*acZ*-BAC transgene was able to direct expression in extraembryonic mesoderm (Chapter III) and deletion of 220 bp ECR2 from the 5' GFP/*acZ*-BAC also resulted in partial loss of extraembryonic mesoderm expression, we obtained transgenic embryos at 7.5 dpc to look specifically for extraembryonic mesoderm expression as well. Similar to the 5' GFP/*acZ*-BAC (FIGURE 7.3b), the 467 bp ECR2-βglobin/*acZ* transgene was sufficient to direct reporter expression in extraembryonic mesoderm in multiple independent embryos generated at 7.5 dpc (FIGURE 7.3b). In addition, two embryos were generated from the 668 bp ECR2-Hsp68/*acZ* transgene (uses the Hsp68 promoter instead of the βglobin promoter described above) at 8.5 dpc. Both embryos exhibited *lacZ* expression in lateral plate mesoderm, although one embryo was highly mosaic (data not shown).

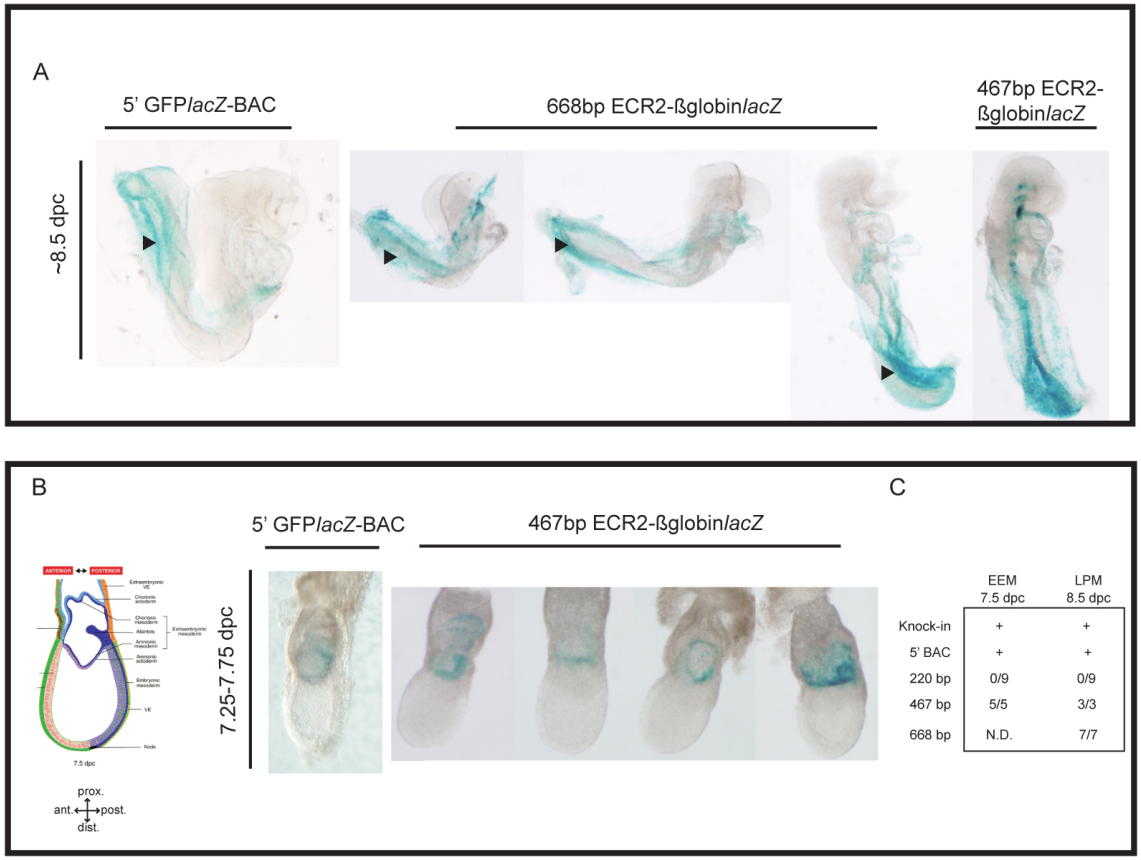

Figure 7.3. ECR2 fragments exhibit mesoderm enhancer activity in transient transgenic mouse embryos. (a) Xgal-stained ECR2-βglobin/*lacZ* transgenic embryos carrying either the 668 bp ECR2 fragment and the 467 bp ECR2 fragment exhibit lateral plate mesoderm expression (arrowheads) similar to lateral plate mesoderm expression seen in 5' GFP/*lacZ*-BAC embryos at 8.5 dpc. (b) Xgal-stained ECR2-βglobin/*lacZ* transient transgenic embryos from the 467 bp ECR2 fragment show lacZ expression in extraembryonic mesoderm similar to extraembryonic mesoderm expression in seen in 5' GFP/*lacZ*-BAC embryos at 7.5 dpc. Note, there is variation in the exact developmental stage in early embryos. (EEM= extraembryonic mesoderm, LPM= Lateral plate mesoderm)

TRANSFAC® Analysis Reveals Putative Transcription Factor Binding Motifs in ECR2

Enhancer elements often contain multiple transcription factor binding sites to allow a combination of transcription factors to bind the DNA and elicit transcription or repress transcription of the target gene (Carey and Smale 2000). To search ECR2 for putative transcription factor binding motifs, the weight matrix-based MATCH™ tool from TRANSFAC® was utilized (Kel et al. 2003) (Matys et al. 2006). This analysis was performed on the three fragments spanning ECR2 that were tested in mouse (see above) and the two fragments spanning ECR2 sequences that were tested in zebrafish (Chapter VI). This allowed us to compare putative transcription factor binding motifs in mouse versus zebrafish sequences, as well as successively larger fragments containing ECR2. Initial analysis focused on minimizing false positive results (see Methods) to reveal binding motifs that have the highest likelihood of matching consensus sequences.

The smallest mouse fragment containing ECR2 (220 bp) had a single Pax6 binding motif, whereas the orthologous zebrafish fragment (249 bp) contained eight transcription factor binding motifs including a motif for Pax6 (FIGURE 7.4). Pax6 is not required for early mesoderm development, nor has *Pax6* expression in early mesoderm been reported (Lang et al. 2007). Interestingly, the 249 bp zebrafish fragment contains a Cdx1 binding motif (FIGURE 7.4, asterisk). *Cdx1* is expressed in both zebrafish and mouse mesoderm (Shimizu et al. 2005) (Meyer and Gruss 1993) during early

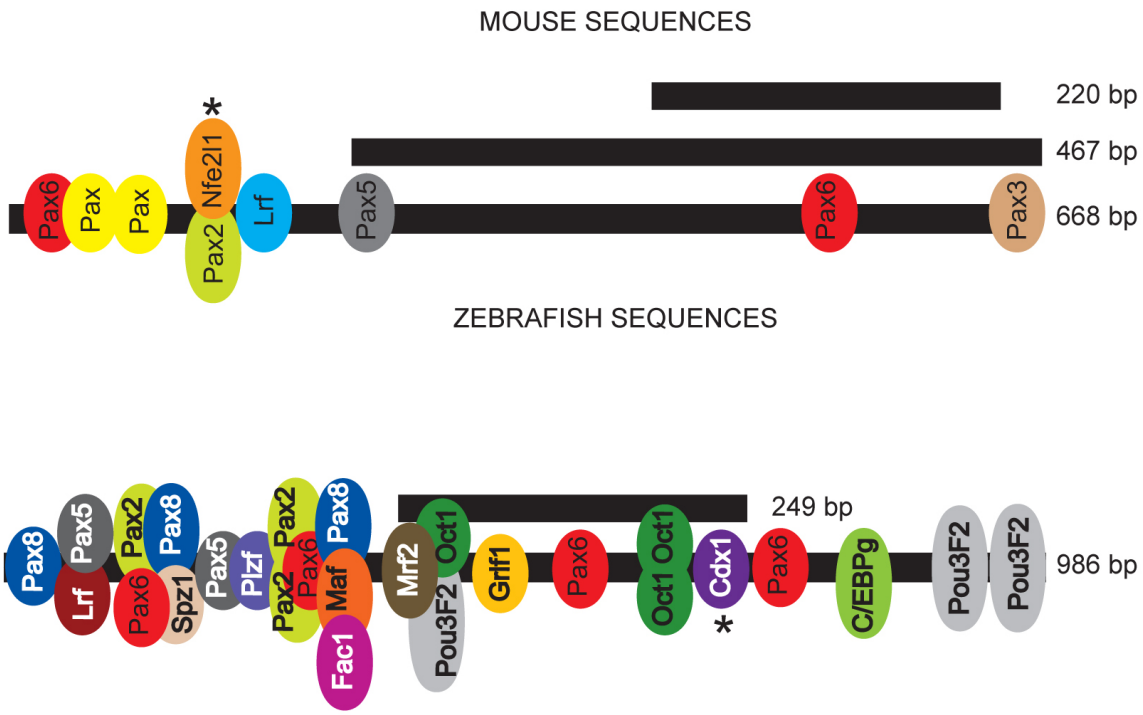

Figure 7.4. TRANSFAC® analysis using a profile to minimize the false positive rate reveals limited putative binding motifs in mouse ECR2 sequences and one binding motif for a transcription factor that is required for mesoderm development. Mouse ECR2 fragments (220, 467, 668 bp) and zebrafish ECR2 fragments (249, 986 bp) are depicted as solid black bars (not drawn to scale). Transcription factors are annotated in their approximate position of the putative binding motif along the lengh of the largest ECR2 fragment. Transcription factors that are expressed in mesoderm are marked with an asterisk (*).

development, although it is not required for early mesoderm development (Subramanian et al. 1995).

As previously stated, the 467 bp mouse fragment tested *in vivo* exhibited enhancer activity in extraembryonic as well as lateral plate mesoderm (FIGURE 7.3). Initial TRANSFAC® analysis of this fragment revealed two additional Pax binding motifs (Pax5, Pax3) in comparison to the 220 bp mouse fragment, neither of which are known to be expressed in mesoderm or required for early mesoderm development (Lang et al. 2007). The largest mouse fragment tested in vivo (668 bp) displayed enhancer activity in lateral plate mesoderm in early mouse embryos (8.5 dpc) (FIGURE 7.3a and c). TRANSFAC® analysis revealed four additional Pax binding motifs as well as a single Lrf binding motif (FIGURE 7.4). It is not known whether the Pax genes or Lrf are expressed in mesoderm. However, neither appears to be required for early mesoderm development (Lang et al. 2007) (Maeda et al. 2007). In addition, a single Nfe2l1/Lcrf1 binding site was predicted in the 668 bp sequence (FIGURE 7.4). Interestingly, both embryonic and extraembryonic mesoderm formation is ablated in Lcrf1-null embryos suggesting Nfe2l1/Lcrf1 is absolutely essential for mesoderm development (Farmer et al. 1997). Finally, TRANSFAC® analysis of the largest zebrafish ECR2 fragment tested in vivo predicted many additional binding motifs (FIGURE 7.4), although none of the transcription binding factors appear to be required for mesoderm development.

Initial TRANSFAC® analysis was geared towards minimizing false positive results and, as a result, very few binding motifs were predicted in relation to the

size of each DNA fragment. In addition, few binding motifs of factors that are critical for mesoderm development (Nfe2l1/Lcrf1) or expressed in mesoderm (Cdx1) were predicted in the mouse and zebrafish sequences, respectively. A single binding motif for Nfe2l1/Lcrf1 was predicted in the 668 bp sequence, as previously stated. However, the 467 and 668 bp sequences each directed mesoderm expression (FIGURE 7.3), suggesting binding motifs for factors expressed in mesoderm or required for mesoderm formation would be present in the sequence common to both fragments tested (467 bp, 668bp). Therefore, additional TRANSFAC® analysis was performed on the mouse sequences alone using alternative parameters to determine if more binding motifs for mesoderm-specific factors could be identified. The profile next utilized was “vertebrate non-redundant minimize the sum of both error rates” and the cutoff selection for the profile was “minimize the sum of both false negative and false positive error rates” (minSUM). We hypothesized that this profile and cutoff selection (minSUM) may identify other binding motifs that were eliminated by the stringent parameters of the initial analysis (minFP) (see Methods). This analysis revealed significantly more transcription factor binding motifs in the three mouse sequences containing ECR2 (220bp=50 binding sites, 467 bp=107 binding sites, 668 bp=178 binding sites).

Next, a gene expression data query for in situ hybridization or in situ reporter results depicting genes expressed in mesoderm or extraembryonic mesoderm during the developmental window when these structures first appear (6.25-8.0 dpc) was performed using the Mouse Genome Informatics database

(http://www.informatics.jax.org/) (Eppig et al. 2005) (Hill et al. 2004). The resulting list of genes expressed in mesoderm/extraembryonic mesoderm (n=215) was compared with the list of predicted binding motifs in one of the three mouse sequences containing ECR2 to identify transcription factors that are predicted to potentially bind the fragment and have expression domains that overlap the mesoderm enhancer's expression (FIGURE 7.3). Interestingly, this analysis predicted multiple mesoderm-specific binding motifs common to both sequences (467 bp, 668 bp) that also directed mesoderm-specific expression, as well as multiple mesoderm-specific binding motifs in the smallest fragment tested (220 bp) (FIGURE 7.5). Eighteen predicted binding motifs for genes expressed in mesoderm were found in the largest ECR2-containing sequence (668 bp), representing six factors expressed in mesoderm (FIGURE 7.5). The smallest ECR2 fragment (220 bp) contained 6/18 predicted mesoderm-specific binding motifs, while the 467 bp fragment contained 11/17 predicted mesoderm-specific binding motifs (FIGURE 7.5).

In comparison to our initial TRANSFAC® analysis using different parameters, this analysis predicted an additional Nfe2l1/Lcrf1 binding site in the sequence shared by all three fragments tested in vivo (FIGURE 7.5). Therefore, this analysis predicted a binding motif for a factor that is required for mesoderm development (Farmer et al. 1997) and is present in both sequences capable of directing mesoderm-specific expression (FIGURE 7.3). In addition to Nfe2l1/ Lcrf1, our analysis predicted all three ECR2-containing sequences contained binding motifs for Cdx1, Zic3, Gata4 and Hand1:E47. *Cdx1* is expressed in

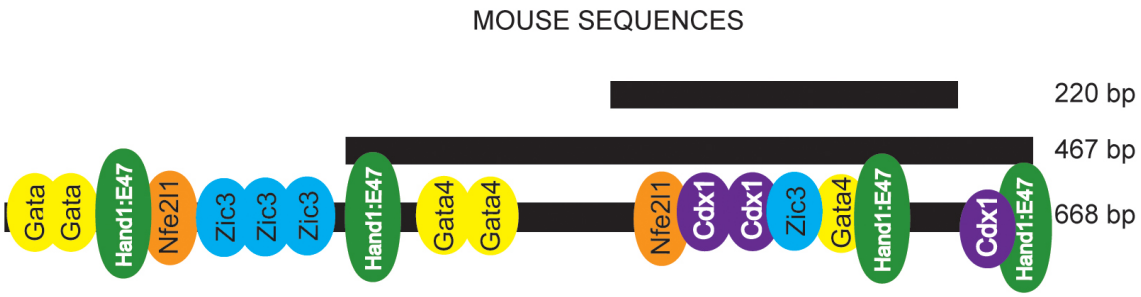

Figure 7.5. TRANSFAC® analysis using a profile to minimize the sum of both error rates identifies numerous binding motifs in mouse ECR2 sequences that are expressed in mesoderm as identified using the Mouse Genome Informatics database query. Mouse ECR2 fragments (220, 467, 668 bp) are represented by solid black bars (not drawn to scale). Transcription factors that are expressed in mesoderm are depicted along the length of the largest ECR2 fragment in the approximate position of the predicted binding motif. Binding factors expressed in mesoderm or extraembryonic mesoderm at 6.25-8.0 dpc in mouse embryos were identified using a query for *in situ* hybridization or *lacZ* reporter results in the Mouse Genome Informatics database.

mouse mesoderm during early development (Meyer and Gruss 1993), although it is not required for early mesoderm development (Subramanian et al. 1995). Zic3 is expressed in embryonic mesoderm and primitive streak, but not in extraembryonic mesoderm (Elms et al. 2004). *Zic3*-null embryos exhibit variable phenotypes, the majority of which involve gastrulation defects (Ware et al. 2006). More specifically, *Zic3*-null Type I mutants fail to develop mesoderm and Type II mutants fail to pattern the primitive streak (Ware et al. 2006). *Gata4* is expressed in mesoderm at 7.5 dpc (Saga et al. 1999) and approximately 33% of *Gata4*-null embryos fail to gastrulate (Molkentin et al. 1997). Finally, *Hand1* is expressed in extraembryonic mesoderm and, later in development, in lateral plate mesoderm (Cserjesi et al. 1995). *Hand1*-null embryos exhibit defects in extraembryonic mesoderm (Firulli et al. 1998). Taken together, four of six mesoderm-specific transcription factors with predicted binding motifs in our ECR2-containing sequences are required for gastrulation or extraembryonic mesoderm formation. These findings are extremely interesting given the mesoderm-specific expression in two of three ECR2-containing fragments (FIGURE 7.3c), as well as the requirement of *Bmp4* for mesoderm development (Winnier et al. 1995).

To visualize the extent of conservation present in mesoderm-specific binding motifs, binding sites were annotated on the Mulan-generated text file showing the alignment of all three ECR2-containing sequences in multiple species (FIGURE 7.6). Mesoderm-specific binding motifs are denoted below the sequence alignment and are also indicated by red font in the 668 bp ECR2-containing mouse sequence (FIGURE 7.6). Note, the position of the 467 bp and

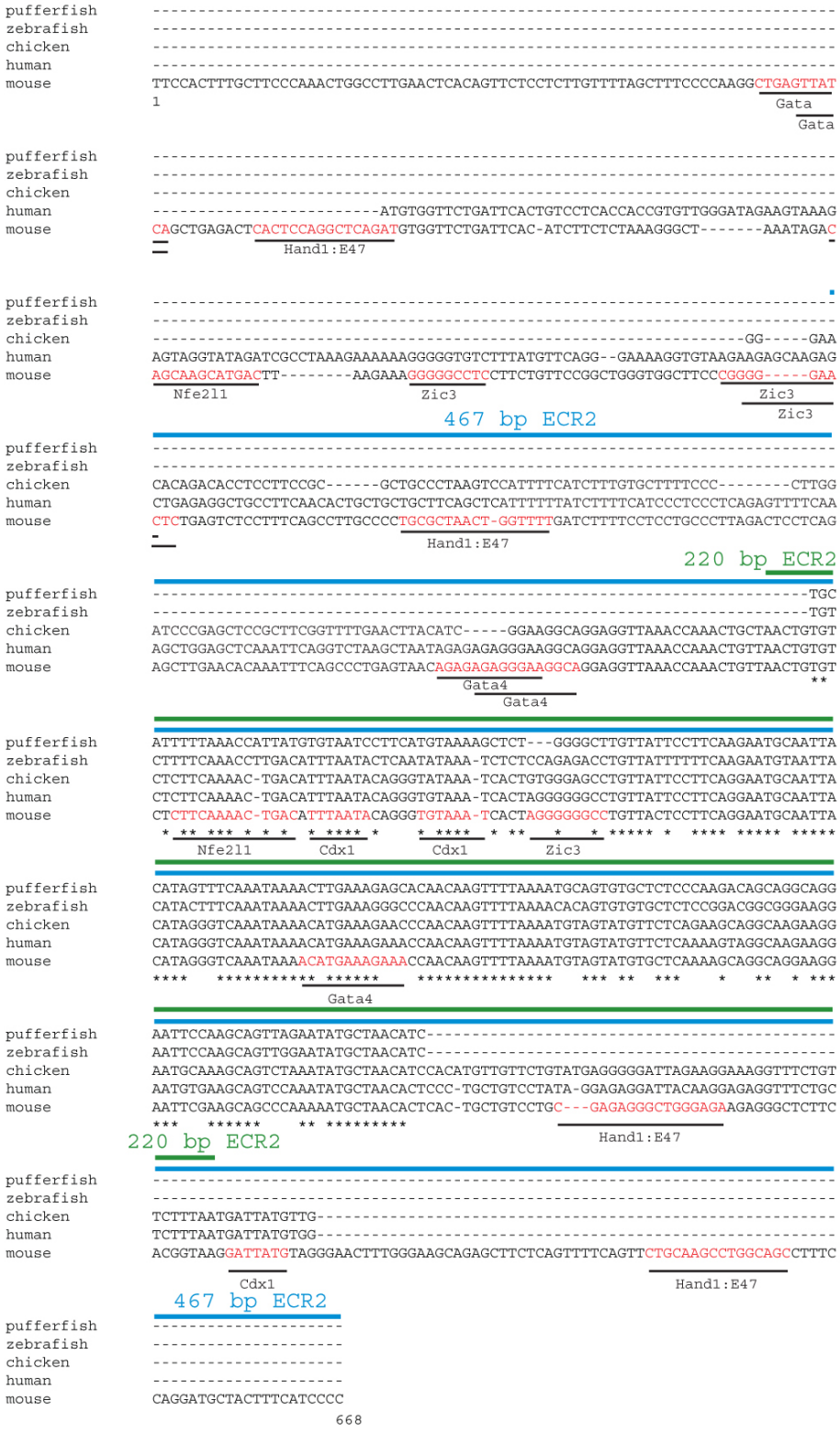

Figure 7.6. Multiple sequence alignment of ECR2-containing sequences (220, 467, 668 bp) depicting binding motifs of transcription factors expressed in mesoderm.

220 bp ECR2-containing sequences relative to the 668 bp sequence are demarcated by blue and green bars, respectively. Mesoderm-specific binding motifs are present in the ancient core conserved sequence (FIGURE 7.6). In addition, binding motifs for Nfe2l1, Hand1, Zic3, Gata4, and Cdx1 are conserved in human, chicken, and mouse (FIGURE 7.6). Interestingly, only human and mouse sequences could be aligned to the region outside the 467 bp fragment but within the 668 bp fragment. This region contains predicted binding motifs for Nfe2l1and Zic3 (FIGURE 7.6).

Discussion

Previous research has indicated the minimal mouse *Bmp4* promoter fragments fail to recapitulate most sites of embryonic *Bmp4* expression in transgenic mouse assays (Feng et al. 2002) (Zhang et al. 2002). In contrast, we have shown BAC-based constructs containing *Bmp4* are sufficient to direct many patterns of expression that reflect endogenous *Bmp4* expression (Chapter III). The complex nature of *Bmp4* expression throughout development (Chapter I) coupled with the presence of numerous conserved noncoding elements dispersed throughout the gene desert encompassing *Bmp4* (Chapter II) suggested many of these noncoding elements could be functional enhancers. Additional comparative sequence analyses revealed three ancient ECRs flanking *Bmp4* (Chapter II) leading us to hypothesize that they are *Bmp4 cis*-regulatory elements. In fact, deletion of an ancient, noncoding, conserved sequence (ECR2) located approximately 50 kb 5' to *Bmp4* resulted in the specific loss of

expression in mesoderm suggesting it is required for *Bmp4* expression (Chapter V). Although deletion of ECR1 and 3 did not reveal an obvious requirement for *Bmp4* expression, transient analysis of all three ECRs in fish suggested 2/3 ECRs exhibited enhancer activity (Chapter VI). Therefore, we tested each ECR for enhancer activity in transgenic mouse reporter assays.

Initial tests of all three ECR-βglobin/*lacZ* constructs failed to demonstrate enhancer activity in multiple independent transgenic mouse lines. However, the sequences that were tested represented conservation between mouse and pufferfish and did not include the larger, multi-vertebrate conserved flanking sequence. Therefore, critical sequences necessary for enhancer function in mouse reside beyond the confines of the pufferfish/mouse conservation. Although mouse/pufferfish conservation of noncoding sequences has proven to be a beacon for identifying enhancer elements, our results strongly suggest it can be advantageous to test a much larger fragment containing the ECR in enhancer assays, rather than testing the minimally conserved sequence.

We showed two larger sequences containing ECR2 (467 bp, 668 bp) displayed mesoderm-specific enhancer activity when tested with the minimal βglobin promoter. The 467 bp ECR2-βglobin/*lacZ* transgene was able to direct extraembryonic mesoderm expression at ~7.5 dpc as well as lateral plate mesoderm expression at ~8.5 dpc, whereas the 220 bp ECR2-βglobin/*lacZ* transgene failed to direct any mesoderm expression. However, the same 220 bp deletion from the 5' GFP/*lacZ*-BAC transgene abolished lateral plate mesoderm expression (Chapter V). Taken together, this suggests there are critical binding

sites that reside in the additional sequence provided by the 467 bp fragment. We cannot rule out the possibility that the 467 bp fragment may actually represent two enhancer modules: extraembryonic mesoderm and lateral plate mesoderm. Although the 220 bp deletion resulted in the complete loss of lateral plate mesoderm expression, partial expression in extraembryonic mesoderm was present (Chapter V)(data not shown). This may be due to remaining functional enhancer-like sequences flanking the 220 bp deletion.

Although the transient analysis of larger ECR2-βglobin*lacZ* transgenes allowed us to quickly test fragments for enhancer activity in mice, this also limited access to essentially two developmental stages (~7.25-8.5 dpc). In addition, the developmental stage of embryos from the same litter can vary, and gene expression in gastrulating embryos changes rapidly causing the traditional "dpc" method of staging to be somewhat inaccurate (Downs and Davies 1993). This is the most likely explanation for subtle variations in reporter expression exhibited in the transient transgenic embryos (FIGURE 7.3). Establishing stable transgenic lines would allow us to methodically age-match gastrulating embryos carrying the same transgene.

MATCH™ analysis is a method that highlights potential transcription factor binding motifs in a sequence, yet it does not functionally test the binding motifs. Thus, it is a hypothesis-generating tool. When utilizing a profile that minimized false positive results, we found a limited number of transcription factor binding motifs in ECR2 sequences. Although this profile is useful for identifying binding motifs that closely match the consensus sequence, it prevented the identification

of non-consensus sequences that could bind to the factor in vivo. The profile that minimized the sum of false positives and negatives allows more binding motifs to be identified. In this regard, numerous potential binding motifs of transcription factors that are expressed in extraembryonic or embryonic mesoderm could be identified in the ECR2-containing sequences. Enhancers have been shown typically to bind a combination of multiple transcription factors to elicit a proper transcriptional response (Carey and Smale 2000). Since the smallest fragment is not sufficient to direct reporter expression in mouse, but is required for reporter expression, we hypothesize a combination of sites present in the 220 bp and the 467 bp sequences work cooperatively to elicit *Bmp4* transcription in mesoderm. Future studies testing the functional significance of putative binding sites will allow researchers to understand what combination of factors binds ECR2 resulting in *Bmp4* transcription.

Taken together, our results indicate a 467 bp noncoding DNA sequence is sufficient to function in a context-independent manner as a *Bmp4 cis*-regulatory element. To our knowledge this is the first tissue-specific *Bmp4* enhancer identified apart from the few kb near the transcription start site. In addition, this *cis*-regulatory element is long-range and functions nearly 50 kb 5' to the minimal *Bmp4* promoter. The significance of this ancient, long-range *Bmp4* mesoderm enhancer is increased by the knowledge that *Bmp4*-null mice fail to develop mesoderm and, as a result, fail to complete embryogenesis (Winnier et al. 1995).

CHAPTER VIII

SUMMARY AND FUTURE DIRECTIONS

A combinatorial approach utilizing comparative analyses and transgenic model organisms has enabled us to begin to tackle the daunting task of understanding how the developmentally crucial gene, *Bmp4*, is transcriptionally regulated. Although previous studies alluded to the possibility that Bmp4 may employ long-range regulatory mechanisms to control its dynamically regulated spatiotemporal expression throughout development (Feng et al. 2002) (Zhang et al. 2002) (Shentu et al. 2003), no data has been published to support this hypothesis until now.

ECR Synteny

To identify putative *cis*-regulatory elements flanking *Bmp4*, comparative analyses of pufferfish and mouse genomic sequences were performed as described in Chapter II. We hypothesized noncoding pufferfish/mouse ECRs would be functional since they had been maintained over 450 million years of evolution. Comparative analyses of pufferfish and mouse identified three noncoding ECRs that have been maintained in the same order and orientation relative to *Bmp4* in multiple vertebrate species to this day. As the genomes of additional species are sequenced and made available through the UCSC Genome Browser, it would be interesting to verify the presence of each ECR in other vertebrate species.

ECR2 Binding Motif Predictions

Two of three ECRs exhibited reporter activity in transient transgenic fish assays discussed in Chapter VI. Although comparative analysis revealed evolutionarily conserved transcription factor binding motifs in all three ECRs, conserved binding factor motifs obtained from rVISTA analysis using the default parameters did not provide clues as to the function of these ECRs. For example, rVISTA analysis showed ECR2 contained conserved binding motifs for transcription factors important in neural development. However, the function of ECR2 is unrelated to neural development since it is a mesoderm enhancer. Therefore, we caution the use of rVISTA (with the default parameters) to formulate hypotheses regarding the function of ECRs. MATCH™ analysis, however, predicted numerous binding motifs in ECR2 for transcription factors that are expressed in mesoderm including several factors that are required for mesoderm development. Although we have not done studies to show that these binding motifs are functional in ECR2, it appears that the MATCH™ analysis provided more promising binding motifs that are in line with the function of ECR2.

Evolution of *Bmp4* Expression in Craniofacial Structures

Identification of *Bmp4* regulatory elements may provide insight into the evolutionary processes that literally shape different species. For example, evolutionary developmental biology studies have addressed the impact of differential *Bmp4* expression on beak size, shape and strength in Darwin's finches (Abzhanov et al. 2004). This study took advantage of six species of

Darwin's finches from the genus Geospiza that are categorized by their beak appearance that bear species-specific distinctions (Abzhanov et al. 2004). Examination of *Bmp4* expression in developing finch beaks revealed distinct spatiotemporal patterns of expression between finches (Abzhanov et al. 2004). Interestingly, expression analysis of genes known to modulate Bmp4 protein expression (Shh, Fgf) had similar expression profiles between finches strongly suggesting the differences in *Bmp4* expression are due to differences in the cis-regulatory control over *Bmp4* expression in the beak (Abzhanov and Tabin 2004) (Abzhanov et al. 2004). In their studies, *Bmp4* expression was modulated in beak mesenchyme and prenasal cartilages. Our studies indicate the *cis*-regulatory element(s) for craniofacial mesenchyme resides in the +25 to +199 kb 3' BAC interval (Chapter III). Identification of this *cis*-regulatory element(s) in mouse could provide a way to identify the homologous element in Darwin's finches by using comparative analyses to identify the *cis*-regulatory element in chicken. Once the *cis*-regulatory element is identified in Darwin's finches, detailed analysis could be performed to determine species-specific sequence differences in the *cis*-regulatory element. This, in turn, may unveil species-specific differential binding of transcription factors allowing distinct modulation of *Bmp4* expression, thus distinct beak shape. Likewise, similar studies have hypothesized a *Bmp4 cis*-regulatory element is responsible for differences in *Bmp4* expression in the African cichlid mandible that leads to species-specific differences in jaw shape (Albertson et al. 2005). Taken together, future studies that identify *cis*-regulatory elements for structures that vary in a species-specific

manner may shed light into the evolutionary processes that give species their morphological adaptations.

Bmp4 Regulatory Landscape Beyond the 400 kb Assayed

To assay the regulatory landscape that contains mouse *Bmp4*, two *Bmp4* GFP*lacZ*-BACs were tested *in vivo* for reporter activity (Chapter III). In contrast to previous reports where minimal *Bmp4* promoter fragments directed a limited number of *Bmp4* expression patterns (Feng et al. 2002) (Zhang et al. 2002), the reporter BACs directed many *Bmp4* expression patterns. Interestingly, many sites of expression were directed by either the 5' or the 3' *Bmp4* GFP*lacZ*-BAC indicating there are numerous long-range *cis*-regulatory elements in addition to the proximal *cis*-regulatory elements that reside in this 400 kb segment encompassing *Bmp4*. However, not all *Bmp4* expression patterns were directed by the 5' or the 3' *Bmp4* GFP*lacZ*-BACs suggesting additional *cis*-regulatory elements reside beyond the 400 kb segment that was tested. In support of this hypothesis, comparative analyses revealed a significant number of noncoding ECRS exist in the extensive 3' desert (Chapter II) beyond the 199 kb 3' segment that was tested. Future studies could focus on testing additional BACs that contain the 3' desert. This would necessitate a slightly different approach since the *Bmp4* transcription unit would not be present in a more distant BAC. Instead, the BAC could be coinjected with a heterologous promoter/reporter construct such as Hsp68 (DiLeone et al. 2000). Upon pronuclear injection, the BAC and Hsp68*lacZ* construct would ligate and form concatamers allowing *cis*-regulatory

elements in the BAC to engage and upregulate the Hsp68*lacZ* construct. This may allow the identification of additional *Bmp4* regulatory elements such as extraembryonic ectoderm, dorsal retina, or anterior limb bud. Alternatively, the 3' GFP*lacZ*-BAC could be ligated to a more distant 3' BAC with a minimal overlapping segment to test additional 3' sequence in the context of the *Bmp4* promoter (Kotzamanis and Huxley 2004).

ECR Deletions Revisited

In Chapter V, we deleted each pufferfish/mouse ECR from the 5' or 3' *Bmp4* GFP*lacZ*-BAC and tested the deletion BACs *in vivo*. Although deletion BAC 1 and 3 failed to demonstrate a requirement for ECR1 or 3 to direct tissue-specific *Bmp4* expression, deletion BAC 2 showed ECR2 was required for mesoderm expression. Future experiments could focus on the potential requirement of ECR1 and 3 by generating transgenic embryos from deletion BAC 1 and deletion BAC 3 lines at additional early embryonic timepoints. For example, embryos were generated at 9.5, 12.5 and 15.5 dpc and stained with Xgal to detect *lacZ* activity. However, it is possible that ECR1 and/or ECR3 direct expression in a transient tissue. If this were true, our analysis may have missed the loss of expression since we assayed three fixed timepoints during embryogenesis. Alternatively, larger deletions could be engineered to remove the core pufferfish/mouse ECR as well as the multi-vertebrate conserved ECR in case the original deletion did not remove all functional ECR sequence. As discussed in Chapter II, other more distant pufferfish/mouse ECRs (ECR4, 5, and

6) were identified by comparative analysis (TABLE 2.1). If more distant 5' and/or 3' GFP/*lacZ*-BACs were tested, deletions of these ECRs could also be tested *in vivo*. Although focusing our efforts on understanding the function of pufferfish/ mouse ECRs has proven to be fruitful, many more mouse/human ECRs are dispersed throughout the *Bmp4* gene desert (Chapter II). To address the potential function of these ECRs with limited constructs, it would be advantageous to generate multiple 5' and 3' GFP/*lacZ*-BACs with large, sequential deletions similar to the approach used to identify *Bmp2* enhancers (Chandler et al. 2007).

Bmp4 Regulatory Architecture and ChIP on Chip

Another approach to decode the regulatory architecture encompassing *Bmp4* would be to employ chromatin immunoprecipitation (ChIP) on chip (Buck and Lieb 2004) (Negre et al. 2006) to assess the location of chromatin-associated factors throughout the *Bmp4* locus. Recent data suggests ChIP can be applied to cell populations as small as 100 cells by using fly chromatin as a carrier substance (O'Neill et al. 2006). By taking advantage of GFP expression in the 5' and 3' GFP/*lacZ*-BACs, GFP-positive cells from transgenic mouse embryos could be sorted by flow cytometry and chemically crosslinked to adhere bound factors to the DNA. Antibodies against the transcription factor of interest, such as Nfe2l1, would be used to immunoprecipitate Nfe2l1-DNA complexes. Next, the DNA that was bound to Nfe2l1 would be purified, amplified, and fluorescently tagged. The immunoprecipitated and fluorescently tagged DNA would be

hybridized to a microchip tiled with oligonucleotides that are complementary to sequences within the genomic interval containing *Bmp4*. The end result would highlight specific locations in the *Bmp4* desert where Nfe2l1 is bound to genomic DNA in the GFP-positive cells.

Testing Enhancer Activity in Fish and Mouse

Transient transgenic methods in zebrafish were used to test the enhancer activity of ECR1, 2 and 3 in Chapter VI. Using this method, our data suggests two of three ECRs have enhancer activity. ECR 2 and 3 displayed reporter activity in the notochord and muscle, respectively. However, whole mount in situ hybridization on zebrafish revealed *Bmp4* is not endogenously expressed in these structures. Since transient transgenic analysis results in highly mosaic expression, it is possible that an ECR directed expression in a transient or small structure and was missed in this type of analysis. Therefore, future studies should focus on establishing stable transgenic lines for each ECR in zebrafish. Stable lines would overcome the mosaicism present in transient transgenic fish allowing an entire structure to express GFP. In addition, passing the transgene through the germline may be required for expression that mimics endogenous *Bmp4*. Although the mouse sequence containing ECR2 directed mesoderm-specific expression, neither fish sequence directed GFP expression in mesoderm during early development (6-10 hpf). However, both fish sequences directed GFP expression in the mesodermally-derived notochord at 24 hpf suggesting these sequences may be capable of directing GFP expression in mesoderm. It is

possible that the transient analysis is not able to detect GFP expression in early embryogenesis (6-10 hpf). Alternatively, it is possible that elements in the proximal *Bmp4* promoter are required for ECR activity. In this case, it would be advantageous to test each ECR in the context of the *Bmp4* promoter. Finally, because little is known about the regulation of *Bmp4* in fish (Shentu et al. 2003), it would be beneficial to generate stable *Bmp4* GFP*lacZ*-BAC lines in fish. These lines could be used to verify whether an ECR-directed site of expression is endogenous and located within the BAC interval.

In Chapter VII, pufferfish/mouse ECRs were tested for their ability to engage and upregulate a heterologous promoter/reporter construct in mouse. No reproducible patterns of expression were evident in any of the ECR-βglobin*lacZ* lines. However, when the ECRs were tested in fish, they exhibited enhancer activity. Since we designed ECR-βglobin*lacZ* constructs to test the core-conserved sequence, we concluded that the pufferfish/mouse ECR sequences were too small and not sufficient to direct reporter expression. Results from experiments testing ECR2 fragments that were twice as large as the original sequence tested and were sufficient to direct tissue-specific reporter expression substantiate this conclusion. Future experiments should test the entire region of multi-vertebrate conservation containing the core pufferfish/ mouse ECR1 or ECR3 in *cis* with the βglobin*lacZ* construct.

It is interesting to note that the smaller sequences containing ECR 2 and 3 showed enhancer activity in fish, but not in mouse. In the context of predicted binding motif data in the larger versus the smaller sequences containing ECR2

(FIGURE 7.6), it is possible that fish only require the smallest sequence for enhancer activity, while mouse requires at least 467 bp. More specifically, the DNA containing predicted binding motifs for Hand1, Gata4 and Cdx1 may be required in conjunction with the smaller sequence containing predicted binding motifs for Hand1, Nfe2l1, Cdx1, Zic3 and Gata4, for enhancer activity in mouse, but not in fish. This hypothesis is consistent with the results from Chapter V indicating the loss of the latter binding motifs ablates expression in lateral plate mesoderm.

Transient analysis of the 467 bp ECR2-βglobin*lacZ* embryos demonstrated ECR2 was sufficient to direct extraembryonic mesoderm expression at ~7.5 dpc and lateral plate mesoderm expression at ~8.5 dpc. Although deletion BAC 2 lines showed the core-conserved ECR2 sequence is required for BAC-directed reporter expression in mesoderm, it does not answer the question of whether ECR2 is required for endogenous expression of *Bmp4* in mesoderm. Therefore, it would be interesting to generate an ECR2-null mouse to test the requirement for ECR2 to direct *Bmp4* expression in mesoderm. Others have shown that disruption of a *Shh* enhancer in mouse phenocopies the congenital defect observed in humans (Lettice et al. 2002). The most severely affected *Bmp4*-null mice fail to develop mesoderm (Winnier et al. 1995). Therefore, by generating *ECR2*-null mice, we could determine if the loss of ECR2 phenocopies *Bmp4*-null mice.

Functional Analysis of Predicted Binding Motifs for Mesoderm-Specific Transcription Factors

MATCH™ analysis on ECR2-containing sequences, suggested the 467 bp ECR2 fragment contains multiple putative binding sites for transcription factors that are expressed in mesoderm. Interestingly, multiple putative binding motifs are for transcription factors that have been shown to be required for normal mesoderm development (Nfe2l1/Lcrf1, Hand1, Gata4, Zic3) (Farmer et al. 1997) (Firulli et al. 1998) (Molkentin et al. 1997) (Ware et al. 2006). Some obvious questions arise from this analysis such as: 1) Are mesodermally-expressed transcription factor binding motifs in ECR2 functional? 2) If a binding motif is functional, do point mutations in the binding site result in the loss or perturbation of mesoderm expression *in vivo*? These questions can be addressed by first performing ChIP on chip experiments, as discussed earlier, to further suggest a transcription factor binds to ECR2 *in vivo*. In addition, electrophorectic mobility shift assays (EMSAs) and *in vitro* footprinting using *in vitro* transcribed and translated protein can be used to determine if candidate factors are capable of binding the putative binding motifs in question. If these tests identify physiologically relevant transcription factor/ECR2 interactions, then transgenes containing point mutations can be tested for reporter activity *in vivo*. These data would suggest whether or not loss of binding site function results in loss or perturbation of ECR2-directed reporter expression. Ultimately, single point mutations of functional binding motifs that are required for reporter expression could be engineered in mouse ES cells to definitively show a point mutation results in the loss of ECR2 function by disrupting transcription factor binding.

ECR2 and Mesoderm Development

Focusing future efforts on ECR2 may provide insight into the process of mesoderm induction (Kimelman 2006). In fact, very few mesoderm-specific enhancers have been studied in great detail (Kimelman 2006). Not only may understanding what upstream components are involved in ECR2's function help researchers to understand mesoderm development, but ECR2 may also be used as a Cre driver to dissect mesoderm development. Cre technology has enabled scientists to delete genes in a cell-specific manner (Branda and Dymecki 2004). For example, ECR2 could be used to drive Cre in mesoderm resulting in the specific deletion of a mesodermally-expressed gene flanked by loxP sites from mesoderm. Or, Cre-loxP could be used to investigate the fate of ECR2/*Bmp4* -expressing mesoderm cells. Taken together, ECR2 may be a useful tool for dissecting the molecular intricacies in early embryogenesis.

ECR2 and Human Disease

Recent data has linked mutations in *BMP4* to developmental defects in human (Bakrania et al. 2008). Understanding the functional role of ECR2 in mouse may reveal a clinically significant role for ECR2 in humans. For instance, *Bmp4* expression in epiblast-derived tissues (extraembryonic mesoderm, lateral plate mesoderm) is required for primordial germ cell survival, allantois and blood vessel development, and normal left-right patterning (Fujiwara et al. 2002). Furthermore, normal left-right patterning is important for cardiac development/ heart looping, as demonstrated in *Bmp4*-null embryos with mesocardia (failure of

the heart to loop) (Fujiwara et al. 2002). Scientists have hypothesized *Bmp4* expression in extraembryonic mesoderm initiates Nodal expression in left-right patterning, while *Bmp4* expression in lateral plate mesoderm maintains *Nodal* expression in left-right patterning (Fujiwara et al. 2002). Therefore, it is possible that point mutations in ECR2 may lead to defects in mesodermal derivatives or defects in downstream signaling cascades. Taken together, human patients with phenotypic defects in mesodermal derivatives or defects in tissues that require Bmp4 signaling in mesoderm for proper development could be screened for mutations in ECR2.

In sum, the research presented in this disseration may provide the momentum needed to further dissect and understand the upstream mechanisms that regulate *Bmp4* expression. Since *Bmp4* is developmentally regulated and dynamically expressed, understanding how it is regulated could impact research on embryogenesis and/or organogenesis. Likewise, studying the conservation and function of an ancient enhancer such as ECR2 may provide answers to evolutionary questions regarding morphological adaptations and DNA sequence function. Finally, this research highlights the complex and important question, how do you define a gene?

REFERENCES

Abzhanov, A., M. Protas, B.R. Grant, P.R. Grant, and C.J. Tabin. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science 305: 1462-5.

Abzhanov, A. and C.J. Tabin. 2004. Shh and Fgf8 act synergistically to drive cartilage outgrowth during cranial development. Dev Biol 273: 134-48.

Ahituv, N., E.M. Rubin, and M.A. Nobrega. 2004. Exploiting human--fish genome comparisons for deciphering gene regulation. Hum Mol Genet 13 Spec No 2: R261-6.

Albertson, R.C., J.T. Streelman, T.D. Kocher, and P.C. Yelick. 2005. Integration and evolution of the cichlid mandible: the molecular basis of alternate feeding strategies. Proc Natl Acad Sci U S A 102: 16287-92.

Alexander, G.M., K.L. Erwin, N. Byers, J.S. Deitch, B.J. Augelli, E.P. Blankenhorn, and T.D. Heiman-Patterson. 2004. Effect of transgene copy number on survival in the G93A SOD1 transgenic mouse model of ALS. Brain Res Mol Brain Res 130: 7-15.

Allende, M.L., M. Manzanares, J.J. Tena, C.G. Feijoo, and J.L. Gomez-Skarmeta. 2006. Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. Methods 39: 212-9.

Aono, A., M. Hazama, K. Notoya, S. Taketomi, H. Yamasaki, R. Tsukuda, S. Sasaki, and Y. Fujisawa. 1995. Potent ectopic bone-inducing activity of bone morphogenetic protein-4/7 heterodimer. Biochem Biophys Res Commun 210: 670-7.

Aparicio, S., A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf, and S. Brenner. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc Natl Acad Sci U S A 92: 1684-8.

Bakrania, P., M. Efthymiou, J.C. Klein, A. Salt, D.J. Bunyan, A. Wyatt, C.P. Ponting, A. Martin, S. Williams, V. Lindley, J. Gilmore, M. Restori, A.G. Robson, M.M. Neveu, G.E. Holder, J.R. Collin, D.O. Robinson, P. Farndon, H. Johansen-Berg, D. Gerrelli, and N.K. Ragge. 2008. Mutations in BMP4 cause eye, brain, and digit developmental anomalies: overlap between the BMP4 and hedgehog signaling pathways. Am J Hum Genet 82: 304-19.

Baleato, R.M., R.J. Aitken, and S.D. Roman. 2005. Vitamin A regulation of BMP4 expression in the male germ line. Dev Biol 286: 78-90.

Balemans, W. and W. Van Hul. 2002. Extracellular regulation of BMP signaling in vertebrates: a cocktail of modulators. Dev Biol 250: 231-50.

Ballester, M., A. Castello, E. Ibanez, A. Sanchez, and J.M. Folch. 2004. Real-time quantitative PCR-based system for determining transgene copy number in transgenic animals. Biotechniques 37: 610-3.

Barton, L.M., B. Gottgens, M. Gering, J.G. Gilbert, D. Grafham, J. Rogers, D. Bentley, R. Patient, and A.R. Green. 2001. Regulation of the stem cell leukemia (SCL) gene: a tale of two fishes. Proc Natl Acad Sci U S A 98: 6747-52.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. Science 304: 1321-5.

Bell, A.C., A.G. West, G. Felsenfeld. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98: 387-96.

Beppu, H., M. Kawabata, T. Hamamoto, A. Chytil, O. Minowa, T. Noda, and K. Miyazono. 2000. BMP type II receptor is required for gastrulation and early development of mouse embryos. Dev Biol 221: 249-58.

Bishop, J. 1996. Chromosomal insertion of foreign DNA. Reprod Nutr Dev. 36: 607-18.

Bishop, J.O. and P. Smith. 1989. Mechanism of chromosomal integration of microinjected DNA. Mol Biol Med 6: 283-98.

Bitgood, M.J. and A.P. McMahon. 1995. Hedgehog and Bmp genes are coexpressed at many diverse sites of cell-cell interaction in the mouse embryo. Dev Biol 172: 126-38.

Blackman, R.K., R. Grimaila, M.M. Koehler, and W.M. Gelbart. 1987. Mobilization of hobo elements residing within the decapentaplegic gene complex: suggestion of a new hybrid dysgenesis system in Drosophila melanogaster. Cell 49: 497-505.

Blackman, R.K., M. Sanicola, L.A. Raftery, T. Gillevet, and W.M. Gelbart. 1991. An extensive 3' cis-regulatory region directs the imaginal disk expression of decapentaplegic, a member of the TGF-beta family in Drosophila. Development 111: 657-66.

Boffelli, D., M.A. Nobrega, and E.M. Rubin. 2004. Comparative genomics at the vertebrate extremes. Nat Rev Genet 5: 456-65.

Bondarenko, V.A., Y.V. Liu, Y.I. Jiang, and V.M. Studitsky. 2003. Communication over a large distance: enhancers and insulators. Biochem Cell Biol 81: 241-51.

Branda, C.S. and S.M. Dymecki. 2004. Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. Dev Cell 6: 7-28.

Brandt, W, M. Khandekar, N. Suzuki, M. Yamamoto, K.C. Lim, and J.D. Engel. Defining the functional boundaries of the gata2 locus by rescue with a linked bacterial artificial chromosome transgene. J Biol Chem 283: 8976-83.

Buck, M.J. and J.D. Lieb. 2004. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83: 349-60.

Burkhart, C.A., M.D. Norris, and M. Haber. 2002. A simple method for the isolation of genomic DNA from mouse tail free of real-time PCR inhibitors. J Biochem Biophys Methods 52: 145-9.

Canalis, E., A.N. Economides, and E. Gazzerro. 2003. Bone morphogenetic proteins, their antagonists, and the skeleton. Endocr Rev 24: 218-35.

Canestro, C., H. Yokoi, and J.H. Postlethwait. 2007. Evolutionary developmental biology and genomics. Nat Rev Genet 8: 932-42.

Carey, M. and S.T. Smale. 2000. Transcriptional Regulation in Eukaryotes. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Carninci, P. et al. 2005. The transcriptional landscape of the mammalian genome. Science 309: 1559-63.

Carroll, S.B., Grenier, Jennifer K., Weatherbee, Scott D. 2001. From DNA to Diversity. Blackwell Science, Malden, Massachusetts.

Chandler, K.J., R.L. Chandler, E.M. Broeckelmann, Y. Hou, E.M. Southard-Smith, and D.P. Mortlock. Relevance of BAC transgene copy number in mice: transgene copy number variation across multiple transgenic lines and correlations with transgene integrity and expression. Mamm Genome 18: 693-708.

Chandler, R.L., K.J. Chandler, K.A. McFarland, and D.P. Mortlock. 2007. Bmp2 transcription in osteoblast progenitors is regulated by a distant 3' enhancer located 156.3 kilobases from the promoter. Mol Cell Biol. 27: 2934-51.

Chang, H., D. Huylebroeck, K. Verschueren, Q. Guo, M.M. Matzuk, and A. Zwijsen. 1999. Smad5 knockout mice die at mid-gestation due to multiple embryonic and extraembryonic defects. Development 126: 1631-42.

Chang, H. and M.M. Matzuk. 2001. Smad5 is required for mouse primordial germ cell development. Mech Dev 104: 61-7.

Charles, M.A., T.L. Saunders, W.M. Wood, K. Owens, A.F. Parlow, S.A. Camper, E.C. Ridgway, and D.F. Gordon. 2006. Pituitary-specific Gata2 knockout: effects on gonadotrope and thyrotrope function. Mol Endocrinol 20: 1366-77.

Chen, D., M. Zhao, S.E. Harris, and Z. Mi. 2004. Signal transduction and biological functions of bone morphogenetic proteins. Front Biosci 9: 349-58.

Chocron, S., M.C. Verhoeven, F. Rentzsch, M. Hammerschmidt, and J. Bakkers. 2007. Zebrafish Bmp4 regulates left-right asymmetry at two distinct developmental time points. Dev Biol 305: 577-88.

Constam, D.B. and E.J. Robertson. 1999. Regulation of bone morphogenetic protein activity by pro domains and proprotein convertases. J Cell Biol 144: 139-49.

Cook, M.J. 1965. The Anatomy of a Laboratory Mouse. Academic Press.

Copeland, N.G., N.A. Jenkins, and D.L. Court. 2001. Recombineering: a powerful new tool for mouse functional genomics. Nat Rev Genet 2: 769-79.

Coucouvanis, E. and G.R. Martin. 1999. BMP signaling plays a role in visceral endoderm differentiation and cavitation in the early mouse embryo. Development 126: 535-46.

Cserjesi, P., D. Brown, G.E. Lyons, and E.N. Olson. 1995. Expression of the novel basic helix-loop-helix gene eHAND in neural crest derivatives and extraembryonic membranes during mouse development. Dev Biol 170: 664-78.

Cui, Y., R. Hackenmiller, L. Berg, F. Jean, T. Nakayama, G. Thomas, and J.L. Christian. 2001. The activity and signaling range of mature BMP-4 is regulated by sequential cleavage at two sites within the prodomain of the precursor. Genes Dev 15: 2797-802.

Cui, Y., F. Jean, G. Thomas, and J.L. Christian. 1998. BMP-4 is proteolytically activated by furin and/or PC6 during vertebrate embryonic development. Embo J 17: 4735-43.

Davidson, E.H. 2001. Genomic Regulatory Systems. Academic Press, San Diego.

Deal, K.K., V.A. Cantrell, R.L. Chandler, T.L. Saunders, D.P. Mortlock, and E.M. Southard-Smith. 2006. Distant regulatory elements in a Sox10-beta GEO BAC transgene are required for expression of Sox10 in the enteric nervous system and other neural crest-derived tissues. Dev Dyn 235: 1413-32.

Dermitzakis, E.T., A. Reymond, and S.E. Antonarakis. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet 6: 151-7.

Dick, A., A. Meier, and M. Hammerschmidt. 1999. Smad1 and Smad5 have distinct roles during dorsoventral patterning of the zebrafish embryo. Dev Dyn 216: 285-98.

DiLeone, R.J., G.A. Marcus, M.D. Johnson, and D.M. Kingsley. 2000. Efficient studies of long-distance Bmp5 gene regulation using bacterial artificial chromosomes. Proc Natl Acad Sci U S A 97: 1612-7.

DiLeone, R.J., L.B. Russell, and D.M. Kingsley. 1998. An extensive 3' regulatory region controls expression of Bmp5 in specific anatomical structures of the mouse embryo. Genetics 148: 401-8.

Downs, K.M. and T. Davies. 1993. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. Development 118: 1255-66.

Dunn, N.R., G.E. Winnier, L.K. Hargett, J.J. Schrick, A.B. Fogo, and B.L. Hogan. 1997. Haploinsufficient phenotypes in Bmp4 heterozygous null mice and modification by mutations in Gli3 and Alx4. Dev Biol 188: 235-47.

Ebara, S., S. Kawasaki, I. Nakamura, T. Tsutsumimoto, K. Nakayama, T. Nikaido, and K. Takaoka. 1997. Transcriptional regulation of the mBMP-4 gene through an E-box in the 5'-flanking promoter region involving USF. Biochem Biophys Res Commun 240: 136-41.

Eblaghie, M.C., M. Reedy, T. Oliver, Y. Mishina, and B.L. Hogan. 2006. Evidence that autocrine signaling through Bmpr1a regulates the proliferation, survival and morphogenetic behavior of distal lung epithelial cells. Dev Biol 291: 67-82.

Elgar, G., R. Sandford, S. Aparicio, A. Macrae, B. Venkatesh, and S. Brenner. 1996. Small is beautiful: comparative genomics with the pufferfish (Fugu rubripes). Trends Genet 12: 145-50.

Elms, P., A. Scurry, J. Davies, C. Willoughby, T. Hacker, D. Bogani, and R. Arkell. 2004. Overlapping and distinct expression domains of Zic2 and Zic3 during mouse gastrulation. Gene Expr Patterns 4: 505-11.

Eppig, J.T., C.J. Bult, J.A. Kadin, J.E. Richardson, J.A. Blake, A. Anagnostopoulos, R.M. Baldarelli, M. Baya, J.S. Beal, S.M. Bello, W.J. Boddy, D.W. Bradt, D.L. Burkart, N.E. Butler, J. Campbell, M.A. Cassell, L.E. Corbani, S.L. Cousins, D.J. Dahmen, H. Dene, A.D. Diehl, H.J. Drabkin, K.S. Frazer, P. Frost, L.H. Glass, C.W. Goldsmith, P.L. Grant, M. Lennon-Pierce, J. Lewis, I. Lu, L.J. Maltais, M. McAndrews-Hill, L. McClellan, D.B. Miers, L.A. Miller, L. Ni, J.E. Ormsby, D. Qi, T.B. Reddy, D.J. Reed, B. Richards-Smith, D.R. Shaw, R. Sinclair, C.L. Smith, P. Szauter, M.B. Walker, D.O. Walton, L.L. Washburn, I.T. Witham, and Y. Zhu. 2005. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res 33: D471-5.

Farmer, S.C., C.W. Sun, G.E. Winnier, B.L. Hogan, and T.M. Townes. 1997. The bZIP transcription factor LCR-F1 is essential for mesoderm formation in mouse development. Genes Dev 11: 786-98.

Feng, J.Q., D. Chen, A.J. Cooney, M.J. Tsai, M.A. Harris, S.Y. Tsai, M. Feng, G.R. Mundy, and S.E. Harris. 1995. The mouse bone morphogenetic protein-4 gene. Analysis of promoter utilization in fetal rat calvarial osteoblasts and regulation by COUP-TFI orphan receptor. J Biol Chem 270: 28364-73.

Feng, J.Q., J. Zhang, X. Tan, Y. Lu, D. Guo, and S.E. Harris. 2002. Identification of cis-DNA regions controlling Bmp4 expression during tooth morphogenesis in vivo. J Dent Res 81: 6-10.

Firulli, A.B., D.G. McFadden, Q. Lin, D. Srivastava, and E.N. Olson. 1998. Heart and extra-embryonic mesodermal defects in mouse embryos lacking the bHLH transcription factor Hand1. Nat Genet 18: 266-70.

Frank, D.B., A. Abtahi, D.J. Yamaguchi, S. Manning, Y. Shyr, A. Pozzi, H.S. Baldwin, J.E. Johnson, and M.P. de Caestecker. 2005. Bone morphogenetic protein 4 promotes pulmonary vascular remodeling in hypoxic pulmonary hypertension. Circ Res 97: 496-504.

Fujiwara, T., D.B. Dehart, K.K. Sulik, and B.L. Hogan. 2002. Distinct requirements for extra-embryonic and embryonic bone morphogenetic protein 4 in the formation of the node and primitive streak and coordination of left-right asymmetry in the mouse. Development 129: 4685-96.

Fujiwara, T., N.R. Dunn, and B.L. Hogan. 2001. Bone morphogenetic protein 4 in the extraembryonic mesoderm is required for allantois development and

the localization and survival of primordial germ cells in the mouse. Proc Natl Acad Sci U S A 98: 13739-44.

Furuta, Y., D.W. Piston, and B.L. Hogan. 1997. Bone morphogenetic proteins (BMPs) as regulators of dorsal forebrain development. Development 124: 2203-12.

Ghanem, N., O. Jarinova, A. Amores, Q. Long, G. Hatch, B.K. Park, J.L. Rubenstein, and M. Ekker. 2003. Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. Genome Res 13: 533-43.

Gilbert, S.F. 2003. Developmental Biology. Sinauer Associates Inc., Sunderland.

Giraldo, P. and L. Montoliu. 2001. Size matters: use of YACs, BACs and PACs in transgenic animals. Transgenic Res 10: 83-103.

Giraldo, P., S. Rival-Gervier, L.M. Houdebine, and L. Montoliu. 2003. The potential benefits of insulators on heterologous constructs in transgenic animals. Transgenic Res 12: 751-5.

Gomez-Skarmeta, J.L., B. Lenhard, and T.S. Becker. 2006. New technologies, new findings, and new concepts in the study of vertebrate cis-regulatory sequences. Dev Dyn 235: 870-85.

Gong, S., C. Zheng, M.L. Doughty, K. Losos, N. Didkovsky, U.B. Schambra, N.J. Nowak, A. Joyner, G. Leblanc, M.E. Hatten, and N. Heintz. 2003. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. Nature 425: 917-25.

Goode, D.K., P. Snell, and G. Elgar. 2003. Comparative analysis of vertebrate Shh genes identifies novel conserved non-coding sequence. Mamm Genome 14: 192-201.

Goode, D.K., P. Snell, S.F. Smith, J.E. Cooke, and G. Elgar. 2005. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. Genomics 86: 172-81.

Groppe, J., J. Greenwald, E. Wiater, J. Rodriguez-Leon, A.N. Economides, W. Kwiatkowski, M. Affolter, W.W. Vale, J.C. Belmonte, and S. Choe. 2002. Structural basis of BMP signalling inhibition by the cystine knot protein Noggin. Nature 420: 636-42.

Hamada, T., H. Sasaki, R. Seki, and Y. Sakaki. 1993. Mechanism of chromosomal integration of transgenes in microinjected mouse eggs:

sequence analysis of genome-transgene and transgene-transgene junctions at two loci. Gene 128: 197-202.

Han, J., M. Ishii, P. Bringas, Jr., R.L. Maas, R.E. Maxson, Jr., and Y. Chai. 2007. Concerted action of Msx1 and Msx2 in regulating cranial neural crest cell differentiation during frontal bone development. Mech Dev 124: 729-45.

Hazama, M., A. Aono, N. Ueno, and Y. Fujisawa. 1995. Efficient expression of a heterodimer of bone morphogenetic protein subunits using a baculovirus expression system. Biochem Biophys Res Commun 209: 859-66.

Heaney, J.D. and S.K. Bronson. 2006. Artificial chromosome-based transgenes in the study of genome function. Mamm Genome 17: 791-807.

Heintz, N. 2001. BAC to the future: the use of bac transgenic mice for neuroscience research. Nat Rev Neurosci 2: 861-70.

Hill, D.P., D.A. Begley, J.H. Finger, T.F. Hayamizu, I.J. McCright, C.M. Smith, J.S. Beal, L.E. Corbani, J.A. Blake, J.T. Eppig, J.A. Kadin, J.E. Richardson, and M. Ringwald. 2004. The mouse Gene Expression Database (GXD): updates and enhancements. Nucleic Acids Res 32: D568-71.

Hogan, B., Beddington, R., Costantini, F., Lacy, E. 1994. Summary of Mouse Development. In Manipulating the Mouse Embryo, pp. 19-114. Cold Spring Harbor Laboratory Press, Plainview.

Hogan, B.L. 1996. Bone morphogenetic proteins: multifunctional regulators of vertebrate development. Genes Dev 10: 1580-94.

Hogan, B.L., R. Beddington, F. Constantini, and E. Lacy. 1994. Maniupulating the Mouse Embryo. Cold Spring Harbor Laboratory Press.

Hua, H., Y.Q. Zhang, S. Dabernat, M. Kritzik, D. Dietz, L. Sterling, and N. Sarvetnick. 2006. BMP4 regulates pancreatic progenitor cell expansion through Id2. J Biol Chem 281: 13574-80.

Huang, J.D., D.H. Schwyter, J.M. Shirokawa, and A.J. Courey. 1993. The interplay between multiple enhancer and silencer elements defines the pattern of decapentaplegic expression. Genes Dev 7: 694-704.

Hubbard, T.J., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S.C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J.

Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. 2007. Ensembl 2007. Nucleic Acids Res 35: D610-7.

Imai, T., R. Takakuwa, S. Marchand, E. Dentz, J.M. Bornert, N. Messaddeq, O. Wendling, M. Mark, B. Desvergne, W. Wahli, P. Chambon, and D. Metzger. 2004. Peroxisome proliferator-activated receptor gamma is required in mature white and brown adipocytes for their survival in the mouse. Proc Natl Acad Sci U S A 101: 4543-7.

Iyengar, A., F. Muller, and N. Maclean. 1996. Regulation and expression of transgenes in fish -- a review. Transgenic Res 5: 147-66.

Jackson, P.D. and F.M. Hoffmann. 1994. Embryonic expression patterns of the Drosophila decapentaplegic gene: separate regulatory elements control blastoderm expression and lateral ectodermal expression. Dev Dyn 199: 28-44.

Jaenisch, R. 1988. Transgenic animals. Science 240: 1468-74.

Jeong, Y., K. El-Jaick, E. Roessler, M. Muenke, and D.J. Epstein. 2006. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. Development 133: 761-72.

Jiao, K., H. Kulessa, K. Tompkins, Y. Zhou, L. Batts, H.S. Baldwin, and B.L. Hogan. 2003. An essential role of Bmp4 in the atrioventricular septation of the mouse heart. Genes Dev 17: 2362-7.

Jones, C.M., K.M. Lyons, and B.L. Hogan. 1991. Involvement of Bone Morphogenetic Protein-4 (BMP-4) and Vgr-1 in morphogenesis and neurogenesis in the mouse. Development 111: 531-42.

Jowett, T. 2001. Double in situ hybridization techniques in zebrafish. Methods 23: 345-58.

Kawakami, K., H. Takeda, N. Kawakami, M. Kobayashi, N. Matsuda, and M. Mishina. 2004. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. Dev Cell 7: 133-44.

Kel, A.E., E. Gossling, I. Reuter, E. Cheremushkin, O.V. Kel-Margoulis, and E. Wingender. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576-9.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12: 656-64.

Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. Genome Res 12: 996-1006.

Kimelman, D. 2006. Mesoderm induction: from caps to chips. Nat Rev Genet 7: 360-72.

Kimura-Yoshida, C., K. Kitajima, I. Oda-Ishii, E. Tian, M. Suzuki, M. Yamamoto, T. Suzuki, M. Kobayashi, S. Aizawa, and I. Matsuo. 2004. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. Development 131: 57-71.

Kinder, S.J., T.E. Tsang, G.A. Quinlan, A.K. Hadjantonakis, A. Nagy, and P.P. Tam. 1999. The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. Development 126: 4691-701.

Kingsley, D.M. 1994. The TGF-beta superfamily: new members, new receptors, and new genetic tests of function in different organisms. Genes Dev 8: 133-46.

Kotzamanis, G. and C. Huxley. 2004. Recombining overlapping BACs into a single larger BAC. BMC Biotechnol 4: 1.

Kurihara, T., K. Kitamura, K. Takaoka, and H. Nakazato. 1993. Murine bone morphogenetic protein-4 gene: existence of multiple promoters and exons for the 5'-untranslated region. Biochem Biophys Res Commun 192: 1049-56.

Lander, E.S. et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Lang, D., S.K. Powell, R.S. Plummer, K.P. Young, and B.A. Ruggeri. 2007. PAX genes: roles in development, pathophysiology, and cancer. Biochem Pharmacol 73: 1-14.

Lawson, K.A., N.R. Dunn, B.A. Roelen, L.M. Zeinstra, A.M. Davis, C.V. Wright, J.P. Korving, and B.L. Hogan. 1999. Bmp4 is required for the generation of primordial germ cells in the mouse embryo. Genes Dev 13: 424-36.

Lawson, K.A., J.J. Meneses, and R.A. Pedersen. 1991. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. Development 113: 891-911.

Lechleider, R.J., J.L. Ryan, L. Garrett, C. Eng, C. Deng, A. Wynshaw-Boris, and A.B. Roberts. 2001. Targeted mutagenesis of Smad1 reveals an essential role in chorioallantoic fusion. Dev Biol 240: 157-67.

Lee, E.C., D. Yu, J. Martinez de Velasco, L. Tessarollo, D.A. Swing, D.L. Court, N.A. Jenkins, and N.G. Copeland. 2001. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. Genomics 73: 56-65.

Lettice, L.A., S.J. Heaney, L.A. Purdie, L. Li, P. de Beer, B.A. Oostra, D. Goode, G. Elgar, R.E. Hill, and E. de Graaff. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12: 1725-35.

Lettice, L.A., T. Horikoshi, S.J. Heaney, M.J. van Baren, H.C. van der Linde, G.J. Breedveld, M. Joosse, N. Akarsu, B.A. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S.W. Scherer, P. Heutink, R.E. Hill, and S. Noji. 2002. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A 99: 7548-53.

Leung, A.Y., E.M. Mendenhall, T.T. Kwan, R. Liang, C. Eckfeldt, E. Chen, M. Hammerschmidt, S. Grindley, S.C. Ekker, and C.M. Verfaillie. 2005. Characterization of expanded intermediate cell mass in zebrafish chordin morphant embryos. Dev Biol 277: 235-54.

Li, Q., S. Harju, and K.R. Peterson. 1999. Locus control regions: coming of age at a decade plus. Trends Genet 15: 403-8.

Li, S., R.E. Hammer, J.B. George-Raizen, K.C. Meyers, and W.T. Garrard. 2000. High-level rearrangement and transcription of yeast artificial chromosome-based mouse Ig kappa transgenes containing distal regions of the contig. J Immunol 164: 812-24.

Liu, W., J. Selever, D. Wang, M.F. Lu, K.A. Moses, R.J. Schwartz, and J.F. Martin. 2004. Bmp4 signaling is required for outflow-tract septation and branchial-arch artery remodeling. Proc Natl Acad Sci U S A 101: 4489-94.

Loots, G.G., R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, and K.A. Frazer. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288: 136-40.

Loots, G.G. and I. Ovcharenko. 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res 32: W217-21.

Lu, C.C., J. Brennan, and E.J. Robertson. 2001. From fertilization to gastrulation: axis formation in the mouse embryo. Curr Opin Genet Dev 11: 384-92.

Maas, S.A. and J.F. Fallon. 2005. Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. Dev Dyn 232: 345-8.

Maconochie, M.K., S. Nonchev, M. Studer, S.K. Chan, H. Popperl, M.H. Sham, R.S. Mann, and R. Krumlauf. 1997. Cross-regulation in the mouse HoxB complex: the expression of Hoxb2 in rhombomere 4 is regulated by Hoxb1. Genes Dev 11: 1885-95.

Maeda, T., T. Merghoub, R.M. Hobbs, L. Dong, M. Maeda, J. Zakrzewski, M.R. van den Brink, A. Zelent, H. Shigematsu, K. Akashi, J. Teruya-Feldstein, G. Cattoretti, and P.P. Pandolfi. 2007. Regulation of B versus T lymphoid lineage fate decision by the proto-oncogene LRF. Science 316: 860-6.

Majewski, I.J., D. Metcalf, L.A. Mielke, D.L. Krebs, S. Ellis, M.R. Carpinelli, S. Mifsud, L. Di Rago, J. Corbin, N.A. Nicola, D.J. Hilton, and W.S. Alexander. 2006. A mutation in the translation initiation codon of Gata-1 disrupts megakaryocyte maturation and causes thrombocytopenia. Proc Natl Acad Sci U S A 103: 14146-51.

Martinez-Barbera, J.P., H. Toresson, S. Da Rocha, and S. Krauss. 1997. Cloning and expression of three members of the zebrafish Bmp family: Bmp2a, Bmp2b and Bmp4. Gene 198: 53-9.

Masucci, J.D. and F.M. Hoffmann. 1993. Identification of two regions from the Drosophila decapentaplegic gene required for embryonic midgut development and larval viability. Dev Biol 159: 276-87.

Masucci, J.D., R.J. Miltenberger, and F.M. Hoffmann. 1990. Pattern-specific expression of the Drosophila decapentaplegic gene in imaginal disks is regulated by 3' cis-regulatory elements. Genes Dev 4: 2011-23.

Matys, V., O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108-10.

Mayor, C., M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, and I. Dubchak. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046-7.

Mercader, N. 2007. Early steps of paired fin development in zebrafish compared with tetrapod limb development. Dev Growth Differ 49: 421-37.

Meyer, B.I. and P. Gruss. 1993. Mouse Cdx-1 expression during gastrulation. Development 117: 191-203.

Mishina, Y. 2003. Function of bone morphogenetic protein signaling during mouse development. Front Biosci 8: d855-69.

Mishina, Y., A. Suzuki, N. Ueno, and R.R. Behringer. 1995. Bmpr encodes a type I bone morphogenetic protein receptor that is essential for gastrulation during mouse embryogenesis. Genes Dev 9: 3027-37.

Miyazaki, Y., K. Oshima, A. Fogo, and I. Ichikawa. 2003. Evidence that bone morphogenetic protein 4 has multiple biological functions during kidney and urinary tract development. Kidney Int 63: 835-44.

Molkentin, J.D., Q. Lin, S.A. Duncan, and E.N. Olson. 1997. Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. Genes Dev 11: 1061-72.

Mortlock, D.P., C. Guenther, and D.M. Kingsley. 2003. A general approach for identifying distant regulatory elements applied to the Gdf6 gene. Genome Res 13: 2069-81.

Muller, F., P. Blader, and U. Strahle. 2002. Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements. Bioessays 24: 564-72.

Muller, F., D.W. Williams, J. Kobolak, L. Gauvry, G. Goldspink, L. Orban, and N. Maclean. 1997. Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. Mol Reprod Dev 47: 404-12.

Nakayama, T., Y. Cui, and J.L. Christian. 2000. Regulation of BMP/Dpp signaling during embryonic development. Cell Mol Life Sci 57: 943-56.

Negre, N., S. Lavrov, J. Hennetin, M. Bellis, and G. Cavalli. 2006. Mapping the distribution of chromatin proteins by ChIP on chip. Methods Enzymol 410: 316-41.

Nikaido, M., M. Tada, T. Saji, and N. Ueno. 1997. Conservation of BMP signaling in zebrafish mesoderm patterning. Mech Dev 61: 75-88.

Nobrega, M.A., I. Ovcharenko, V. Afzal, and E.M. Rubin. 2003. Scanning human gene deserts for long-range enhancers. Science 302: 413.

Nobrega, M.A., Y. Zhu, I. Plajzer-Frick, V. Afzal, and E.M. Rubin. 2004. Megabase deletions of gene deserts result in viable mice. Nature 431: 988-93.

Nomura, M. and E. Li. 1998. Smad2 role in mesoderm formation, left-right patterning and craniofacial development. Nature 393: 786-90.

O'Neill, L.P., M.D. VerMilyea, and B.M. Turner. 2006. Epigenetic characterization of the early embryo with a chromatin immunoprecipitation protocol applicable to small cell populations. Nat Genet 38: 835-41.

Onichtchouk, D., Y.G. Chen, R. Dosch, V. Gawantka, H. Delius, J. Massague, and C. Niehrs. 1999. Silencing of TGF-beta signalling by the pseudoreceptor BAMBI. Nature 401: 480-5.

Ovcharenko, I., G.G. Loots, B.M. Giardine, M. Hou, J. Ma, R.C. Hardison, L. Stubbs, and W. Miller. 2005a. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Res 15: 184-94.

Ovcharenko, I., G.G. Loots, M.A. Nobrega, R.C. Hardison, W. Miller, and L. Stubbs. 2005b. Evolution and functional classification of vertebrate gene deserts. Genome Res 15: 137-45.

Ovcharenko, I., M.A. Nobrega, G.G. Loots, and L. Stubbs. 2004. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. Nucleic Acids Res 32: W280-6.

Padgett, R.W., J.M. Wozney, and W.M. Gelbart. 1993. Human BMP sequences can confer normal dorsal-ventral patterning in the Drosophila embryo. Proc Natl Acad Sci U S A 90: 2905-9.

Palmiter, R.D. and R.L. Brinster. 1986. Germ-line transformation of mice. Annu Rev Genet 20: 465-99.

Pandolfi, P.P., M.E. Roth, A. Karis, M.W. Leonard, E. Dzierzak, F.G. Grosveld, J.D. Engel, and M.H. Lindenbaum. 1995. Targeted disruption of the GATA3 gene causes severe abnormalities in the nervous system and in fetal liver haematopoiesis. Nat Genet 11: 40-4.

Parameswaran, M. and P.P. Tam. 1995. Regionalisation of cell fate and morphogenetic movement of the mesoderm during mouse gastrulation. Dev Genet 17: 16-28.

Pennacchio, L.A., N. Ahituv, A.M. Moses, S. Prabhakar, M.A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K.D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B.L. Black, O. Couronne, M.B. Eisen, A. Visel, and E.M. Rubin. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499-502.

Pennisi, E. 2007. Genetics. Working the (gene count) numbers: finally, a firm answer? Science 316: 1113.

Plessy, C., T. Dickmeis, F. Chalmel, and U. Strahle. 2005. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. Trends Genet 21: 207-10.

Roessler, E., D.E. Ward, K. Gaudenz, E. Belloni, S.W. Scherer, D. Donnai, J. Siegel-Bartelt, L.C. Tsui, and M. Muenke. 1997. Cytogenetic rearrangements involving the loss of the Sonic Hedgehog gene at 7q36 cause holoprosencephaly. Hum Genet 100: 172-81.

Rossi, J.M., N.R. Dunn, B.L. Hogan, and K.S. Zaret. 2001. Distinct mesodermal signals, including BMPs from the septum transversum mesenchyme, are required in combination for hepatogenesis from the endoderm. Genes Dev 15: 1998-2009.

Saga, Y., S. Miyagawa-Tomita, A. Takagi, S. Kitajima, J. Miyazaki, and T. Inoue. 1999. MesP1 is expressed in the heart precursor cells and required for the formation of a single heart tube. Development 126: 3437-47.

Sampath, T.K., K.E. Rashka, J.S. Doctor, R.F. Tucker, and F.M. Hoffmann. 1993. Drosophila transforming growth factor beta superfamily proteins induce endochondral bone formation in mammals. Proc Natl Acad Sci U S A 90: 6004-8.

Sandelin, A., P. Bailey, S. Bruce, P.G. Engstrom, J.M. Klos, W.W. Wasserman, J. Ericson, and B. Lenhard. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5: 99.

Schwartz, S., Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10: 577-86.

Selever, J., W. Liu, M.F. Lu, R.R. Behringer, and J.F. Martin. 2004. Bmp4 in limb bud mesoderm regulates digit pattern by controlling AER development. Dev Biol 276: 268-79.

Shentu, H., H.J. Wen, G.M. Her, C.J. Huang, J.L. Wu, and S.P. Hwang. 2003. Proximal upstream region of zebrafish bone morphogenetic protein 4 promoter directs heart expression of green fluorescent protein. Genesis 37: 103-12.

Shimizu, T., Y.K. Bae, O. Muraoka, and M. Hibi. 2005. Interaction of Wnt and caudal-related genes in zebrafish posterior body formation. Dev Biol 279: 125-41.

Siepel, A., G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.W. Hillier, S. Richards, G.M. Weinstock, R.K. Wilson, R.A. Gibbs, W.J. Kent, W. Miller, and D. Haussler. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-50.

Smit, A., R. Hubley, and P. Green. 1996-2004. RepeatMasker Open-3.0. In.

Solnica-Krezel, L. and W. Driever. 2001. The role of the homeodomain protein Bozozok in zebrafish axis formation. Int J Dev Biol 45: 299-310.

Spencer, F.A., F.M. Hoffmann, and W.M. Gelbart. 1982. Decapentaplegic: a gene complex affecting morphogenesis in Drosophila melanogaster. Cell 28: 451-61.

Spitz, F., F. Gonzalez, and D. Duboule. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113: 405-17.

St Johnston, R.D., F.M. Hoffmann, R.K. Blackman, D. Segal, R. Grimaila, R.W. Padgett, H.A. Irick, and W.M. Gelbart. 1990. Molecular organization of the decapentaplegic gene in Drosophila melanogaster. Genes Dev 4: 1114-27.

Stemple, D.L. 2005. Structure and function of the notochord: an essential organ for chordate development. Development 132: 2503-12.

Stuart, G.W., J.V. McMurray, and M. Westerfield. 1988. Replication, integration and stable germ-line transmission of foreign sequences injected into early zebrafish embryos. Development 103: 403-12.

Stuart, G.W., J.R. Vielkind, J.V. McMurray, and M. Westerfield. 1990. Stable lines of transgenic zebrafish exhibit reproducible patterns of transgene expression. Development 109: 577-84.

Subramanian, V., B.I. Meyer, and P. Gruss. 1995. Disruption of the murine homeobox gene Cdx1 affects axial skeletal identities by altering the mesodermal expression domains of Hox genes. Cell 83: 641-53.

Summerbell, D., P.R. Ashby, O. Coutelle, D. Cox, S. Yee, and P.W. Rigby. 2000. The expression of Myf5 in the developing mouse embryo is controlled by discrete and dispersed enhancers specific for particular populations of skeletal muscle precursors. Development 127: 3745-57.

Szabo, P., S.H. Tang, A. Rentsendorj, G.P. Pfeifer, and J.R. Mann. 2000. Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Curr Biol 10: 607-10.

Tam, P.P. and D.A. Loebel. 2007. Gene function in mouse embryogenesis: get set for gastrulation. Nat Rev Genet 8: 368-81.

Tesson, L., J.M. Heslan, S. Menoret, and I. Anegon. 2002. Rapid and accurate determination of zygosity in transgenic animals by real-time quantitative PCR. Transgenic Res 11: 43-8.

Thermes, V., C. Grabher, F. Ristoratore, F. Bourrat, A. Choulika, J. Wittbrodt, and J.S. Joly. 2002. I-SceI meganuclease mediates highly efficient transgenesis in fish. Mech Dev 118: 91-8.

Thisse, B., V. Heyer, A. Lux, V. Alunni, A. Degrave, I. Seiliez, J. Kirchner, J.P. Parkhill, and C. Thisse. 2004. Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. Methods Cell Biol 77: 505-19.

Thompson, D.L., L.M. Gerlach-Bank, K.F. Barald, and R.J. Koenig. 2003. Retinoic acid repression of bone morphogenetic protein 4 in inner ear development. Mol Cell Biol 23: 2277-86.

Tremblay, K.D., N.R. Dunn, and E.J. Robertson. 2001. Mouse embryos lacking Smad1 signals display defects in extra-embryonic tissues and germ cell formation. Development 128: 3609-21.

Trousse, F., P. Esteve, and P. Bovolenta. 2001. Bmp4 mediates apoptotic cell death in the developing chick eye. J Neurosci 21: 1292-301.

Ureta-Vidal, A., L. Ettwiller, and E. Birney. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4: 251-62.

Urist, M.R. 1965. Bone: formation by autoinduction. Science 150: 893-9.

Van den Wijngaard, A., M.A. Pijpers, P.H. Joosten, J.M. Roelofs, E.J. Van zoelen, and W. Olijve. 1999. Functional characterization of two promoters in the human bone morphogenetic protein-4 gene. J Bone Miner Res 14: 1432-41.

van den Wijngaard, A., M. van Kraay, E.J. van Zoelen, W. Olijve, and C.J. Boersma. 1996. Genomic organization of the human bone morphogenetic protein-4 gene: molecular basis for multiple transcripts. Biochem Biophys Res Commun 219: 789-94.

Venter, J.C. et al. 2001. The sequence of the human genome. Science 291: 1304-51.

Waldrip, W.R., E.K. Bikoff, P.A. Hoodless, J.L. Wrana, and E.J. Robertson. 1998. Smad2 signaling in extraembryonic tissues determines anterior-posterior polarity of the early mouse embryo. Cell 92: 797-808.

Wang, J.M., G.G. Prefontaine, M.E. Lemieux, L. Pope, M.A. Akimenko, and R.J. Hache. 1999. Developmental effects of ectopic expression of the glucocorticoid receptor DNA binding domain are alleviated by an amino acid substitution that interferes with homeodomain binding. Mol Cell Biol 19: 7106-22.

Ware, S.M., K.G. Harutyunyan, and J.W. Belmont. 2006. Zic3 is critical for early embryonic patterning during gastrulation. Dev Dyn 235: 776-85.

Warming, S., N. Costantino, D.L. Court, N.A. Jenkins, and N.G. Copeland. 2005. Simple and highly efficient BAC recombineering using galK selection. Nucleic Acids Res 33: e36.

Warren, A.J., W.H. Colledge, M.B. Carlton, M.J. Evans, A.J. Smith, and T.H. Rabbitts. 1994. The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. Cell 78: 45-57.

Waterston, R.H. et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420: 520-62.

Weinstein, M., X. Yang, C. Li, X. Xu, J. Gotay, and C.X. Deng. 1998. Failure of egg cylinder elongation and mesoderm induction in mouse embryos lacking the tumor suppressor smad2. Proc Natl Acad Sci U S A 95: 9378-83.

Westerfield, M. 2000. The zebrafish book: A guide for the laboratory use of zebrafish (Danio rerio). University of Oregon Press, Eugene.

Whitman, M. and L. Raftery. 2005. TGFbeta signaling at the summit. Development 132: 4205-10.

Winnier, G., M. Blessing, P.A. Labosky, and B.L. Hogan. 1995. Bone morphogenetic protein-4 is required for mesoderm formation and patterning in the mouse. Genes Dev 9: 2105-16.

Wirth, J., T. Wagner, J. Meyer, R.A. Pfeiffer, H.U. Tietze, W. Schempp, and G. Scherer. 1996. Translocation breakpoints in three patients with campomelic dysplasia and autosomal sex reversal map more than 130 kb from SOX9. Hum Genet 97: 186-93.

Woolfe, A., M. Goodson, D.K. Goode, P. Snell, G.K. McEwen, T. Vavouri, S.F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y.J. Edwards, J.E. Cooke, and G. Elgar. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.

Wozney, J.M., V. Rosen, A.J. Celeste, L.M. Mitsock, M.J. Whitters, R.W. Kriz, R.M. Hewick, and E.A. Wang. 1988. Novel regulators of bone formation: molecular clones and activities. Science 242: 1528-34.

Wray, G.A. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8: 206-16.

Wray, G.A., M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, and L.A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377-419.

Wunderle, V.M., R. Critcher, N. Hastie, P.N. Goodfellow, and A. Schedl. 1998. Deletion of long-range regulatory elements upstream of SOX9 causes campomelic dysplasia. Proc Natl Acad Sci U S A 95: 10649-54.

Wurst, W., A.B. Auerbach, and A.L. Joyner. 1994. Multiple developmental defects in Engrailed-1 mutant mice: an early mid-hindbrain deletion and patterning defects in forelimbs and sternum. Development 120: 2065-75.

Zhang, J., X. Tan, C.H. Contag, Y. Lu, D. Guo, S.E. Harris, and J.Q. Feng. 2002. Dissection of promoter control modules that direct Bmp4 expression in the epithelium-derived components of hair follicles. Biochem Biophys Res Commun 293: 1412-9.

Zhao, G.Q. 2003. Consequences of knocking out BMP signaling in the mouse. Genesis 35: 43-56.

Zhu, N.L., C. Li, J. Xiao, and P. Minoo. 2004. NKX2.1 regulates transcription of the gene for human bone morphogenetic protein-4 in lung epithelial cells. Gene 327: 25-36.

Zimmerman, L.B., J.M. De Jesus-Escobar, and R.M. Harland. 1996. The Spemann organizer signal noggin binds and inactivates bone morphogenetic protein 4. Cell 86: 599-606.